Volume 106
Number 4
Winter 2020

Journal of the

WASHINGTON

ACADEMY OF SCIENCES

ISSN 0043-0439

Issued Quarterly at Washington DC

Washington Academy of Sciences

Founded in 1898

Journal of the Washington Academy of Sciences (ISSN 0043-0439)

Published by the Washington Academy of Sciences
email: wasjournal@washacadsci.org
website: www.washacadsci.org

The Journal of the Washington Academy of Sciences

The *Journal* is the official organ of the Academy. It publishes articles on science policy, the history of science, critical reviews, original science research, proceedings of scholarly meetings of its Affiliated Societies, and other items of interest to its members. It is published quarterly. The last issue of the year contains a directory of the current membership of the Academy.

Subscription Rates

Members, fellows, and life members in good standing receive the Journal free of charge. Subscriptions are available on a calendar year basis, payable in advance. Payment must be made in US currency at the following rates.

US and Canada	$30.00
Other Countries	$35.00
Single Copies (when available)	$15.00

Claims for Missing Issues

Claims must be received within 65 days of mailing. Claims will not be allowed if non-delivery was the result of failure to notify the Academy of a change of address.

Notification of Change of Address

Address changes should be sent promptly to the Academy Office. Notification should contain both old and new addresses and zip codes.

Postmaster:

Send address changes to WAS, Rm GL117,
1200 New York Ave. NW
Washington, DC 20005

Academy Office
Washington Academy of Sciences
Room GL117
1200 New York Ave. NW
Washington, DC 20005
Phone: (202) 326-8975

Volume 106
Number 4
Winter 2020

Journal of the

WASHINGTON

ACADEMY OF SCIENCES

ISSN 0043-0439

Issued Quarterly at Washington DC

EDITOR'S COMMENTS

Presenting the 2020 winter issue of the *Journal of the Washington Academy of Sciences*.

There are six papers in this issue plus one interesting Science Bite. All six papers are from an Ontology Conference and comprise a special issue in Ontology.

Please consider submitting short (typically one page) papers on an interesting tidbit in science. There are a lot of interesting tidbits out there. Every science field has them. They sit in your brain ready to share. We all want to learn about things in fields other than our own. So pile them up and send them in.

The Journal is the official organ of the Academy. Please consider sending in technical papers, review studies, announcements, SciBites, and book reviews. Send manuscripts to wasjournal@washacadsci.org. If you are interested in being a reviewer for the *Journal*, please send your name, email address, and specialty to the same address. Each manuscript is peer reviewed, and there are no page charges.

I encourage people to write letters to the editor. Please send by email (wasjournal@washacadsci.org) comments on papers, suggestions for articles, and ideas for what you would like to see in the Journal. I also encourage student papers and will help the student learn about writing a scientific paper.

Please remain safe and healthy in this time of pandemic.

Sethanne Howard

Journal of the Washington Academy of Sciences

Editor Sethanne Howard showard@washacadsci.org

Board of Discipline Editors

The *Journal of the Washington Academy of Sciences* has a twelve member Board of Discipline Editors representing many scientific and technical fields. The members of the Board of Discipline Editors are affiliated with a variety of scientific institutions in the Washington area and beyond — government agencies such as the National Institute of Standards and Technology (NIST); universities such as Georgetown; and professional associations such as the Institute of Electrical and Electronics Engineers (IEEE).

Washington Academy of Sciences
1200 New York Avenue
Rm G119
Washington, DC 20005

Please fill in the blanks and send your application to the address above. We will contact you as soon as your application has been reviewed by the Membership Committee. Thank you for your interest in the Washington Academy of Sciences.

(Dr. Mrs. Mr. Ms)_

Business Address

Home Address

Email

Phone ______

Cell Phone ______

preferred mailing address ____Business _____Home

Type of membership ____Regular ____Student

Schools of Higher Education attended	Degrees	Dates

Present Occupation or Professional Position ______
Please list memberships in scientific societies – include office held

Instructions to Authors

1. Deadlines for quarterly submissions are:

Spring – February 1	Fall – August 1
Summer – May 1	Winter – November 1

2. Draft Manuscripts using a word processing program (such as MSWord), not PDF. We do not accept PDF manuscripts.
3. Papers should be 6,000 words or fewer. If there are seven or more graphics, reduce the number of words by 500 for each graphic.
4. Include an abstract of 150-200 words.
5. Use Times New Roman, font size 12.
6. Include two to three sentence bios of the authors.
7. Graphics must be easily resizable by the editor to fit the Journal's page size. Reference the graphic in the text.
8. Use endnotes or footnotes. The bibliography may be in a style considered standard for the discipline or professional field represented by the paper.
9. Submit papers as email attachments to the editor or to wasjournal@washacadsci.org .
10. Include the author's name, affiliation, and contact information – including postal address. Membership in an Academy-affiliated society may also be noted. It is not required.
11. Manuscripts are peer reviewed and become the property of the Washington Academy of Sciences.
12. There are no page charges.

Washington Academy of Sciences Affiliated Institutions

National Institute for Standards & Technology (NIST)

Meadowlark Botanical Gardens

The John W. Kluge Center of the Library of Congress

Potomac Overlook Regional Park

Koshland Science Museum

American Registry of Pathology

Living Oceans Foundation

National Rural Electric Cooperative Association (NRECA)

Toward Meaningful Explanations

Kenneth Baclawski[1], Mike Bennett[2], Gary Berg-Cross[3],
Todd Schneider[4], Ram D. Sriram[5]

[1]Northeastern University
[2]Hypercube Limited, London
[3]RDA/US Advisory Group, Troy, NY
[4]Engineering Semantics, Fairfax, VA
[5]National Institute of Standards & Technology

Data are important! They are how we understand the world, and understanding the world is the special interest and purpose of Science. Understanding information that we gather about the world is an important part of the scientific process. However, data that are not correctly interpreted and understood are less than useless, they can actually be misleading or even damaging. So how can scientists, and people in general, understand their data? How can they understand the meaning of their data? If someone does not already understand some data, there should be a mechanism whereby an understanding is possible; in other words, some way to explain the data. This special issue is intended for a wide range of people who are concerned with meaningful explanations, including philosophers, physical scientists, engineers, linguists, social scientists, and many others.

SIMPLY PUT AN EXPLANATION is the answer to the question "Why?" as well as the answers to related questions such as "How?" and "Why not?" and requests for details and evidence for an answer. Accordingly, explanations generally occur within the context of a process, which could be a dialog between persons, between a person and a system, or an agent-to-agent communication process between two systems. It is important to note that explanations are not limited to textual media. Visual media such as diagrams, pictures and videos can also express explanations as well as or even better than text, especially when such media are interactive, thereby fulfilling the requirement that explanations allow for subsequent questions and extended conversation. Explanations also occur in social interactions when clarifying a point, expounding a view, or interpreting behavior. Another important context where explanations are important is the process of developing some kind of system, not necessarily a software system. Such a process requires the developers to make a series of decisions. The explanation for a decision is called its decision rationale.

This special issue is devoted to the subject of what explanations are and what they mean. The inspiration for this special issue is the Ontology Summit that was held in the first half of 2019. This event was concerned with the role of applied ontologies for explaining decisions made by a system. While ontology is the branch of philosophy that deals with the nature of being, applied ontology builds on philosophy, cognitive science, linguistics and logic with the purpose of understanding, clarifying, making explicit and communicating people's distinctions and assumptions about the nature and structure of the world. Baclawski *et al* (2019) summarized the findings and challenges that were identified during the Ontology Summit 2019. More specifically, it focused on critical explanation gaps and the role that ontology engineering could play for dealing with these gaps. This special issue expands on the subject of explanations that was introduced by the Ontology Summit 2019.

A brief history of explanations provides some context for this special issue. Among the first known attempts at understanding the why of explanations as explained in (Chatterjee & Dutta, 2014) were those documented among Indian intellectuals and philosophers, beginning with the knowledge collection called the Vedas (dating back to 5000 BCE). This philosophical tradition included notions of context, logic and explanation that are similar to the modern conceptions. For example, there was a notion of syllogism that explicitly incorporated context into the structure of the syllogism. Explanation was also a part of logical inference. More generally, explanation in the form of a dialog between a teacher and a student appears throughout the Vedas (Satprakashananda, 1965; Chennakesavan, 1980).

Greek intellectuals and philosophers subsequently studied the notion of an explanation. For example, to understand and explain the why there was a Peloponnesian War Thucydides defined explanations as a process where facts (indisputable data), which are observed, evaluated based on some common knowledge of human nature. This was then compared in order to reach generalized principles for why some events occur via a process akin to modern induction (Shanske, 2006). In the writings of Plato (*e.g.*, Phaedrus and Theaetetus) we see explanations as an expression using logos knowledge compostable by Universal Forms, which are abstractions of the world's entities we come to experience and know. Facts, in this view, are occurrences or states of affairs and may be a descriptive part of an explanation, but not the deep Why. Aristotle's view, such as in *Posterior Analytics* provides a more familiar view of explanation as part of a logical, deductive, process using reason to reach conclusions. Aristotle proposed four types of causes (αι'τία) to explain things. These were from either the thing's matter, form, end, or change-initiator (efficient cause) (Falcon, 2006). Following Descartes, Leibniz, and especially Newton, modern deterministic causality using natural mechanisms became central to causal explanations. To know what causes an event means to employ natural laws as the central means to understand and explain why it happened. As this makes clear, some notions of the nature of knowledge, namely, how we come to know something and the nature of reality, are parts of explanation. For example, John Stuart Mill provides a deductivist account of explanation as evidenced by these two quotes: "An individual fact is said to be explained, by pointing out its cause, that is by stating the law or laws of causation, of which its production is an instance," and "a law or uniformity of nature is said to be explained, when another law or laws are pointed out, of which that law is but a case, and from which it could be deduced (Mill 1843)."

While explainability has always be a concern of computer systems, the issue has become especially relevant with the success of artificial intelligence (AI) algorithms, such as deep neural networks, whose functioning is too opaque and complex to be understood easily even by those who developed them. This could limit general acceptance of and trust in these algorithms in spite of their advantages and wide range of applicability. Explainable AI (XAI) is an active research area whose goal is to provide AI systems with some degree of explainability. In "Explainable Artificial Intelligence: An Overview," Sargur N. Srihari surveys the field of XAI. Explanations provided by XAI methods take a variety of forms, ranging from traditional feature-based explanations to "heat-map" visualizations, from illustrative examples to probabilistic modeling. Clearly, XAI is an exciting new area at the frontiers of AI.

When computers were developed, one of the earliest questions was whether they might eventually be as intelligent as humans. The field of AI was created not only to investigate this question but also actually to develop systems that achieved it. A fundamental aspect of human intelligence is that we have "common sense," and the study of this aspect of intelligence has been a part of AI from the beginning. AI has also always emphasized the benefits of providing explanations for system reasoning. While commonsense knowledge (CSK) and its associated reasoning processes would seem to be useful for explainability, CSK research has, until recently, been more concerned with knowledge representation than with explainability. In “Commonsense and Explanation: Synergy and Challenges in the Era of Deep Learning Systems” by Gary Berg-Cross, the connections between CSK and explanations are discussed, including the challenges and opportunities. The goal is to achieve fluid explanations that are responsive to changing circumstances, based on commonsense knowledge about the world.

The healthcare enterprise involves many different stakeholders – consumers, healthcare professionals and providers, researchers, and insurers. Sources of health related data are highly diverse and have many levels of granularity. As a result of the COVID-19 pandemic, healthcare issues that were previously only discussed by specialists are now part of the everyday discourse of the average individual. In “Applied Ontologies for Global Health Surveillance and Pandemic Intelligence,” Christopher J. O. Baker, Mohammad Sadnan Al Manir, Jon Hael Brenas, Kate Zinszer, and Arash Shaban-Nejad use Malaria surveillance as a use case to highlight the contribution of applied ontologies for enhancing enhanced interoperability, interpretability and explainability. These technologies are relevant for ongoing pandemic preparedness initiatives.

Financial institutions are very complex entities that play many roles and have many kinds of stakeholders, ranging from customers, to regulators, to shareholders, and to the society as a whole. Given these many responsibilities, it is no surprise that financial institutions “have a lot of explaining to do,” as Michael Bennett so deftly begins his article “Financial Industry Explanation” where he presents some of the challenges of providing meaningful explanation in this domain. Explanations are a special case of the more general requirement of accountability which is becoming an issue for many other domains as well. The lessons learned by the financial industry explainability are likely to be valuable for other domains as well.

Ontologies play a significant role in all of the many research projects referenced by papers in this special issue. However, the ontologies for explainability in XAI, commonsense reasoning, health surveillance, and finance do not seem to have much in common with one another. The final paper, “Decision Rationales as Models for Explanations” by Kenneth Baclawski, attempts to weave the various strands of ontologies for explainability together in a single reference ontology by focusing on the observation that the purpose of most of the systems is to make decisions, and that it is the decisions that need to be explained.

Processes today, whether they are based on software or human activities or a combination of them, or whether they use legacy systems or newly developed systems seldom include explainability. In nearly all cases, explanations are neither recorded nor can be easily generated. Unfortunately, explainability cannot simply be added as another module. Rather it should drive every process from the earliest stages of planning, analysis, and design. Explainability requirements must be empirically discovered during these stages (Clancey 2019). Unfortunately, currently there is little sensitivity to the need for explainability and little experience with addressing it. It is hoped that this special issue will assist stakeholders to develop their systems so that they provide meaningful explanations.

References

Baclawski, K., Bennett, M., Berg-Cross, G., Fritzsche, D., Sharma, R., Singer, J., . . . Whitten, D. (2020).

Ontology Summit 2019 Communiqué: Explanation. Applied Ontology. DOI: 10.3233/AO-200226

Chatterjee, S., & Dutta, D. (2014). An Introduction to Indian Philosophy, Eleventh Impression. Rupa Publications.

Chennakesavan, S. (1980). Concept of mind in indian philosophy. Delhi: Motllal Banarsidass.

Clancey, W. (2019). Explainable AI Past, Present, and Future: A Scientific Modeling Approach. Retrieved on April 28, 2019 from http://bit.ly/2Scjvo6

Falcon, A. (2006). Aristotle on causality. Retrieved 16 September 2020 from https://stanford.io/2ZLknqp

Mill, J. (1843). A system of logic. Harper and Brothers.

Satprakashananda, S. (1965). Methods of Knowledge according to Advaita Vedanta. Advaita Ashram.

Shanske, D. (2006). Thucydides and the philosophical origins of history. Cambridge University Press.

BIO

Kenneth Baclawski is an Associate Professor Emeritus at the College of Computer and Information Science, Northeastern University. Professor Baclawski does research in data semantics, formal methods for software engineering and software modeling, data mining in biology and medicine, semantic collaboration tools, situation awareness, information fusion, self-aware and self-adaptive systems, and wireless communication. He is a member of the Washington Academy of Sciences, IEEE, ACM, IAOA, and is the chair of the Board of Trustees of the Ontolog Forum.

Gary Berg-Cross is a cognitive psychologist (PhD, SUNY–Stony Brook) whose professional life included teaching and R&D in applied data & knowledge engineering, collaboration, and AI research. A board member of the Ontolog Forum he co-chaired the Research Data Alliance work-group on Data Foundations and Terminology. Major thrusts of his work include reusable knowledge, vocabularies, and semantic interoperability achieved through semantic analysis, formalization, capture in knowledge tools, and access through repositories.

Mike Bennett is the director of Hypercube Limited, a company that helps people manage their information assets using formal semantics. Mike is the originator of the Financial Industry Business Ontology (FIBO) from the EDM Council, a formal ontology for financial industry concepts and definitions. Mike provides mentoring and training in the application of formal semantics to business problems and strategy, and is retained as Standards Liaison for the EDM Council and the IOTA Foundation, a novel Blockchain-like ecosystem.

Ram D. Sriram is currently the chief of the Software and Systems Division, Information Technology Laboratory, at the National Institute of Standards and Technology (NIST). Prior to joining NIST, he was on the engineering faculty (1986-1994) at the Massachusetts Institute of Technology (MIT) and was instrumental in setting up the Intelligent Engineering Systems Laboratory. Sriram has co-authored or authored more than 275 publications, and is a Fellow of ASME, AAAS, IEEE and Washington Academy of Sciences, a Distinguished Member (life) of ACM and a Senior Member (life) of AAAI.

Todd Schneider is an ontologist, co-chair of the Industrial Ontology Foundry's Technical Oversight Board, President of Engineering Semantics, Chair of the SCOPE working group, and on the Board of Trustees of the Ontolog Forum. His format training is in physics, mathematics, and mathematical logic. He has developed software and systems large and small and is an expert in interoperability.

Explainable Artificial Intelligence: An Overview

Sargur N. Srihari

University at Buffalo, The State University of New York

Abstract

With a wide range of applications, Artificial Intelligence (AI) has spawned a spectrum of research activity on AI-related topics. One such area is that of explainable AI. It is a vital component of trustworthy AI systems. This paper provides an overview of explainable AI methods describing both post-hoc AI systems, which provide explanations with previously built conventional AI systems, and ante-hoc AI systems, which are configured from the start to provide explanations. The explanations take various forms: explanation based on features, explanation based on illustrative training samples, explanation based on embedded representations, and explanation based on heat-maps. There are also probabilistic explanations which combine neural network models with graphical models. Explainable AI is closely associated with many AI research topic frontiers such as neuro-symbolic AI and machine teaching.

Contents

1. Introduction

ARTIFICIAL INTELLIGENCE (AI) is everywhere. There are billions of searches on handheld devices every day. Smart phones use facial recognition. Alexa cutely answers our questions. The key element in Tik-Tok is its recommender system. More generally, AI enables performing tasks requiring human cognition as well as decision-making.

While AI systems already incorporate intelligent behavior, they continue to be improved with the need to exhibit flexibility, resourcefulness, creativity, real-time responsiveness, and long-term reflection to demonstrate competence in complex environments and social contexts. This paper is an overview of one of the several sub-areas of active AI research known as Explainable AI (XAI). First we set the stage for where XAI fits into the spectrum of AI research topics and methods.

1.1 AI sub-disciplines

While AI has already been incorporated into a wide range of applications, it is also a topic of a great deal of current research. AI research areas can be divided into five areas as follows according to National Science Foundation (2019):

1. *Core AI*: Theory and methods for: (I) learning, abstraction, and inference (II) architectures for intelligence and multi-agent systems. ML has made great advances through algorithms, computing power, and growing data. Other technologies include knowledge representation, logical and probabilistic reasoning, planning, search, constraint satisfaction, and optimization.

2. *Biologically-inspired AI*: Models may be inspired by living systems: connectionism, behavior, and emergence. Computational neuroscience which deals with the theory of computation in the nervous

system Behavioral and cognitive science Typical of human perceptual, motor, and cognitive processes and their interactions

3. *Perception and Communication*: Computer vision methods to sense and reason about visual world. Human language technologies (also called NLP, NLU) to analyze, produce, translate, and respond to human text and speech.

4. *Embodied AI*: Intelligent systems may be able to act upon the world through embodiment. Robotics is closely aligned with but not identical to embodied AI. An embodied AI may be a robot.

5. *Trustworthy AI*: AI amplifies human capabilities to accomplish individual and collective goals. However, there is a need to assess benefits, effects, and risks, as well as how human, technical, and contextual aspects of systems interact to shape those effects. Relevant aspects of trustworthy AI are: Explainable AI (XAI), Validation of AI-enabled systems, AI safety, security, and privacy (including, for example, role of emotion and affect in the design and perception of AI).

1.2 Trustworthy AI

Trust is key in adoption of AI for economic growth and innovations to benefit society. Today, ability to understand AI decisions and measure their trustworthiness is limited. For it to be trustworthy, AI has to be: trusted to function reliably, trusted to be able to explain conclusions, trusted not to violate privacy, and trusted not to exhibit socially harmful bias.

It is the explainability aspect of trustworthy AI that we explore further here. Explanations are vital in decision making. Establishing human trust in the outcome requires the exchange of reasons for that outcome. Explanations must be in terms as appropriate to the task and as needed by users. Explanations help pinpoint errors or data. Research challenges include finding ways to make "black box" AI systems explainable. Models and frameworks for learning and reasoning that are both inherently explainable and powerful. Integrating psychology, cognitive science, to better understand and acceptability of an explanation.

1.3 AI methods

It has long been understood that designing AI needs knowledge. On a historic time-scale, efforts at doing this may be characterized as consisting of several overlapping waves. The first wave of AI began with the knowledge-based approach, where an input is transformed by a hand-designed program to the desired output, *e.g.*, a rule-based expert system. On a parallel track the machine learning approach was developed, where the input is first transformed by a hand-designed program into features but the features were mapped to the output by a program that learns from examples.

The second wave of AI began by replacing feature engineering with representation learning, where the features are learnt automatically. Deep learning involves several representation layers. (Goodfellow, Bengio, and Courville, 2016). First simple features are learnt. Additional layers extract more abstract features. The final layers map the abstract features to the output. Deep learning allows AI systems to rapidly adapt to new tasks, since designing features can take great human effort – often decades for a community of researchers. It does not need programmer to have deep knowledge of the problem domain. The two waves of AI are shown in Fig. 1.

An emerging third wave of AI may be defined as neurosymbolic AI. It is essentially the combination of deep learning with symbolic reasoning. A symbolic reasoning process is used to bridge the learning of visual concepts, words and semantic parsing of sentences without explicit annotations for any of them (Mao, Gan, Kohli, Tenenbaum, and Wu, 2019). Symbolic approaches are usually constructed using graphical models (Koller and Friedman, 2009). Probabilities have been very much a part of the earlier waves, both in discriminative machine learning approaches as well as in generative approaches such as adversarial networks. The neurosymbolic approach relies on generative models at the symbolic level, where the symbols are computed using deep learning. Generative models, such as generative adversarial networks can be used to construct distributions. The symbolic approach calls for algorithms/architectures for: representation (such as Bayesian nets, Markov Random Fields) and inference (Exact, Approximate, Monte Carlo, Variational).

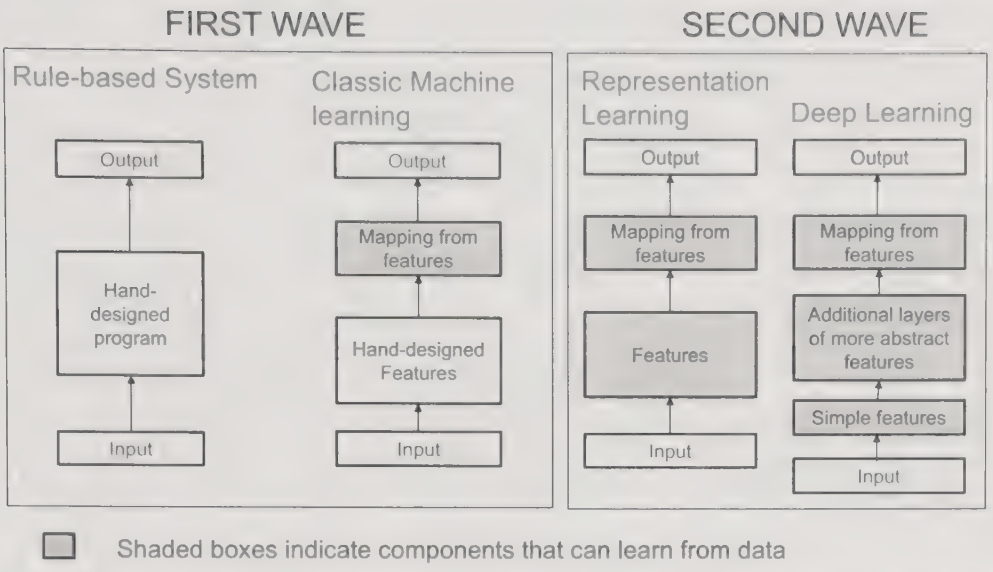


Figure 1: Two waves of AI: (a) The first wave consisted of knowledge-based and machine-learning approaches, and (b) The second wave consists of representation learning and deep learning approaches. Source: (Goodfellow, *et.al*. 2016).

An emerging third wave of AI may be defined as neurosymbolic AI. It is essentially the combination of deep learning with symbolic reasoning. A symbolic reasoning process is used to bridge the learning of visual concepts, words and semantic parsing of sentences without explicit annotations for any of them (Mao, Gan, Kohli, Tenenbaum, and Wu, 2019). Symbolic approaches are usually constructed using graphical models (Koller and Friedman, 2009). Probabilities have been very much a part of the earlier waves, both in discriminative machine learning approaches as well as in generative approaches such as adversarial networks. The neurosymbolic approach relies on generative models at the symbolic level, where the symbols are computed using deep learning. Generative models, such as generative adversarial networks can be used to construct distributions. The symbolic approach calls for algorithms/architectures for: representation (such as Bayesian nets, Markov Random Fields) and inference (Exact, Approximate, Monte Carlo, Variational).

2. Explainable Artificial Intelligence

Explanations are vital in decision making. Establishing human trust in the outcome requires the exchange of reasons for that outcome. Explanations must be in terms as appropriate to the task and as needed by users. Explanations help pinpoint errors or data. Explanations help to detect bias in the system thereby leading to ethical AI.

Research challenges of XAI include finding ways to make "black box" AI systems explainable: Models and frameworks for learning and reasoning that are both inherently explainable and powerful; Integrating psychology, cognitive science, to better understand and acceptability of an explanation.

One of the main criticisms of deep learning is opaqueness, *i.e.*, having the characteristics of a blackbox. Formally, a blackbox is a function that is too complicated for any human to comprehend, a function that is proprietary, or model that is difficult to trouble-shoot. Deep Learning Models are blackbox models because they are recursive, non-intuitive, and difficult for people to understand. See Fig. 2.

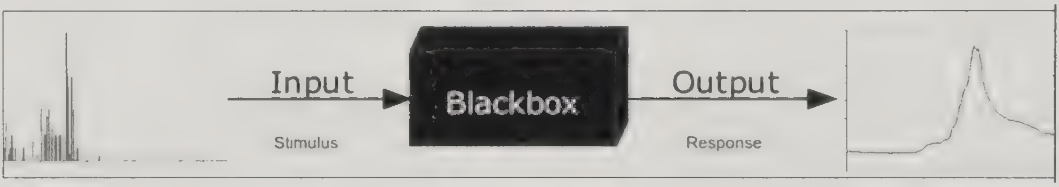


Figure 2: AI as a blackbox. Source: (Wikipedia)

The role of explanation in a human-machine interactive scenario is given in Fig. 3. The explanation interface is one capable of answering the following types of queries, Turek (2018):

1. Why did you do that?
2. Why not something else?
3. When do you succeed?
4. When do you fail?
5. When can I trust you?
6. How do I correct an error?

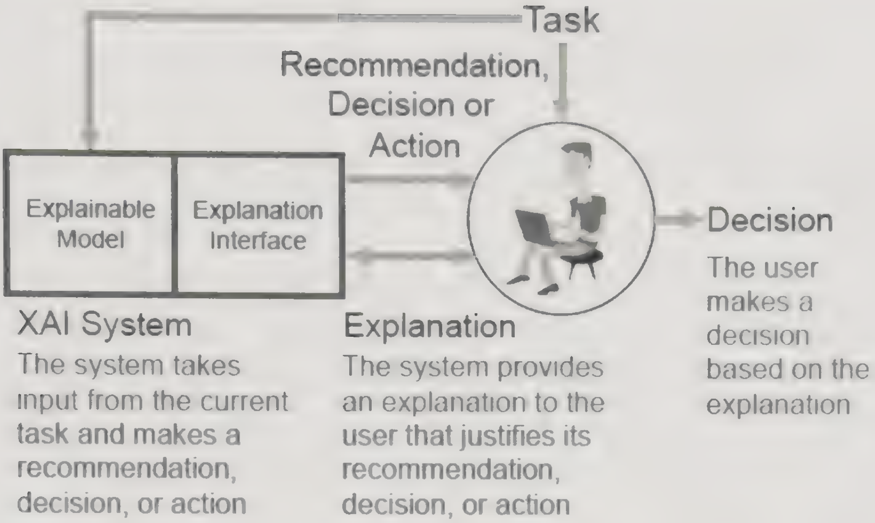

Task
Recommendation, Decision or Action
Explainable Model
Explanation Interface
Decision
The user makes a decision based on the explanation
XAI System
The system takes input from the current task and makes a recommendation, decision, or action
Explanation
The system provides an explanation to the user that justifies its recommendation, decision, or action

Figure 3: Explanation Framework. Source: (Turek, 2018).

2.1 Need for XAI

There are numerous reasons for certain AI deployments to be explainable. Some important ones are: justification, control, discovery, and improvement. We briefly describe each need here.

Explain to justify: The ability to explain one's decision to other people is an important aspect of human intelligence. Understanding the rationale behind the model's predictions would help users decide when to trust or not to trust their predictions.

Explain to control: This refers to compliance to legislation. For instance in making credit decisions, how did the model decide to provide or deny credit to an individual? Was there bias: ethnicity, race religion? In the US the lender must provide reasons for adverse decision, such as take-home insufficient, insufficient collateral, poor credit rating. In the European Union, GDPR (General Data Protection Regulation): right to explanation for high-stakes automated decisions. Another example is healthcare, which is highly regulated due to HIPAA. How did AI predict grade 3 or grade 4 tumor?

Explain to improve: The first step in improving a system is to understand its weaknesses, such as detecting bias in the system. For instance, in the medical domain, an anecdote is that medical AI decisions were worse with AI, *e.g.*, patient discharge to a nursing home did not take into account personal circumstances. An explanation would help the human decision-maker over-ride the system decision.

Explain to discover: AI systems are trained with millions of examples. They may observe previously unseen patterns in data that people may find to be useful.

An example of explanation generation in a deep network in the computer vision domain is given in Fig. 4. Here the goal of the system is to classify the type of bird. After it has generated the class to be a downy woodpecker it also uses the definition of the bird as follows: "This bird has a white breast, black wings and a red spot on its head" to generate the explanation "This is a Downy Woodpecker because it is a black and white bird with a red spot in its crown."

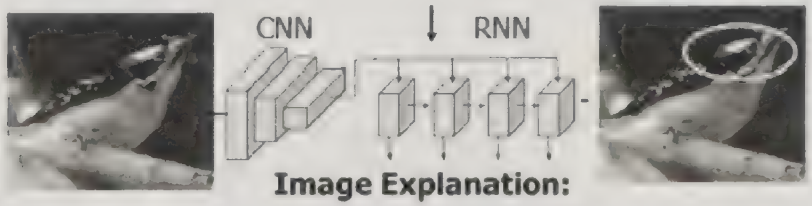


Figure 4: An example of explanation generated by a deep network: "This is a Downy Woodpecker because it is a black and white bird with a red spot in its crown." Source: (Hendricks et al, 2016).

After training a deep network we have large networks that work very well, but hard to tell how. They can fail unintuitively. Adversarial examples show this (See Fig. 5).

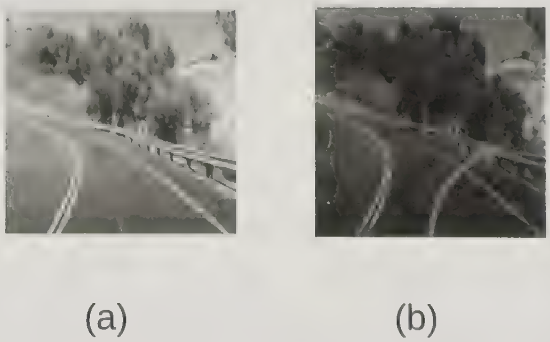


Figure 5: Unexpected failure: (a) correct steering in daytime lighting and (b) wrong steering in fading light. Source image: (Pei, Cao, Yang, and Jana, 2017).

2.2 Measures of explanation effectiveness

Several measures of explanation effectiveness have been proposed Turek (2018):

- User satisfaction: clarity of explanation (user rating), utility of explanation (user rating)
- Mental model: understanding individual decisions, understanding the overall model, strength/weakness assessment, "what will it do" prediction, "How do I intervene" prediction
- Task performance: Does it improve the user's decision, task performance? Artificial decision tasks introduced to diagnose the user's understanding
- Trust assessment: appropriate future use and trust
- Correctability: identifying errors, correcting errors, continuous training

2.3 Taxonomy of XAI methods

With wide adoption of AI in industry and government, the need for XAI has also grown commensurately. Existing XAI methods can be divided into two broad categories (See Fig. 6):

1. Post-hoc (Explain the Blackbox): Explainability based on test cases and results
2. Ante-hoc (Build a new learning model): Seeding explainability into model from the start.

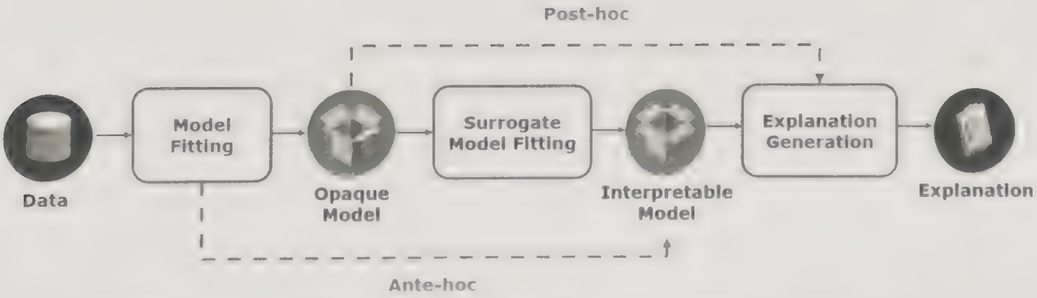


Figure 6: Ante- and Post-Hoc Explainable AI. Source: (Marselis, 2019).

Some examples of post-hoc XAI systems are Sensitivity Analysis (SA), Layer-wise Relevance Propagation (LRP), and Local Interpretable Model-Agnostic Explanations (LIME). Examples of Ante-hoc XAI systems are: Reversed Time Attention Model (RETAIN), and Bayesian Deep Learning (BDL). We discuss each of these types of XAI systems next.

3. Post-hoc XAI

Among methods for visualizing, interpreting and explaining deep learning models, two popular techniques for explaining predictions are Sensitivity Analysis and Layerwise Relevance Propagation. We discuss each of these followed by objectively comparing the quality of the explanations provided.

3.1 Measures of Explanation Quality

Here we describe two measures of explanation quality for evaluating the performance of a deep network: Sensitity Analysis and Layerwise Relevance Propagation. We then compare their efficacy on different tasks.

3.1.1 *Sensitivity Analysis*

Assumes that most relevant features are those to which output is most sensitive. Consider the input image in Fig. 7. The system correctly classifies the input image as “rooster”. Then, an explanation method is applied to explain the prediction in terms of input variables. The result of this explanation process is a heatmap visualizing the importance of each input variable *I* (pixel) for the prediction *f*(*x*). In this example the rooster’s red comb and wattle are the basis for the AI system’s decision. Sensitivity analysis (SA) explains a prediction based on the model’s locally evaluated gradient (partial derivative). This amounts to which pixels need to be changed to make image look more/less like the predicted class, *e.g.*, changing yellow occluding pixels improves score, but does not explain rooster.

How changes in each pixel affect the score are given by the partial derivatives

$$R_I = \left\| \frac{\partial}{\partial x_I} f(x) \right\|$$

SA explains the prediction based on a locally evaluated gradient. However, it does not explain $f(x)$ but a variation of the input.

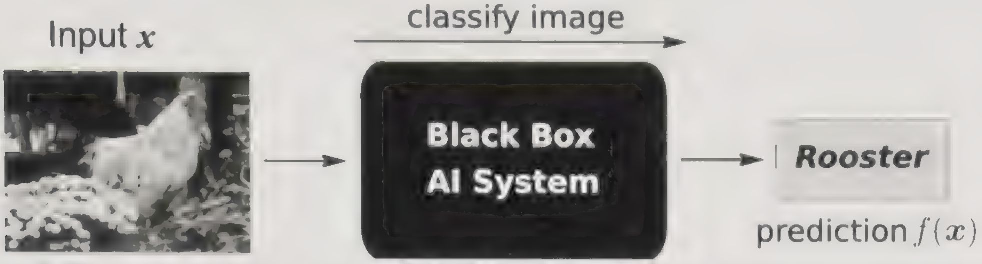


Figure 7: XAI using Sensitivity Analysis: Changing yellow (occluding pixels) improves score, but does not explain rooster. Source: (Samek, Wiegand, and Muller, 2018).

3.1.2 *Layerwise Relevance Propagation*

Layerwise Relevance Propagation (LRP) explains the classifier's prediction using decomposition. See Fig. 8. It redistributes the prediction $f(x)$ backwards using local redistribution rules until it assigns a relevance score R_i to each input variable (*e.g*., image pixel). The key property of this redistribution process is referred to as relevance conservation. It can be summarized by a sequence of sums as follows:

$$\sum R_i = ... = \sum R_j = \sum R_k = f(x)$$

At every step of the redistribution process (*e.g*., at every layer of a deep neural network), the total amount of relevance (*i.e*., the prediction $f(x)$ is conserved. No relevance is artificially added or removed during redistribution. The relevance scores R_i of each input variable determines how much this variable has contributed to the prediction. Thus, in contrast to sensitivity analysis, LRP truly decomposes the function value $f(x)$.

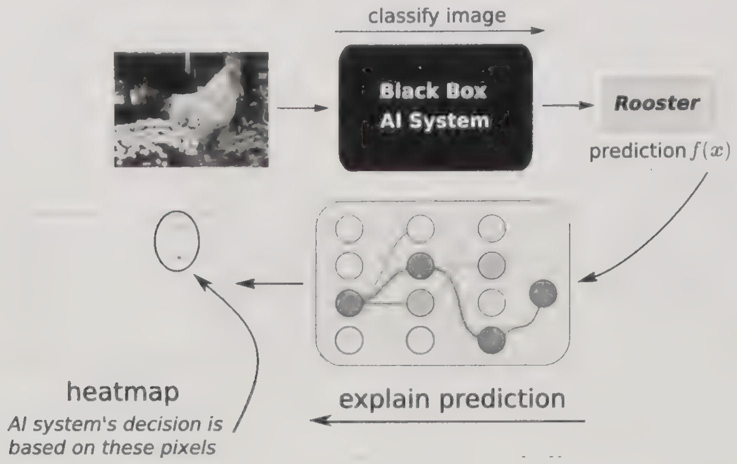


Figure 8: Layerwise Relevance Prediction. In this example the rooster's red comb and wattle are the basis for the AI system's decision. With the heatmap one can verify that the AI system works as intended. Source: (Samek *et al,* 2018).

The LRP redistribution process for feed-forward neural networks is as follows. Let x_j be the neuron activations at layer *l*, R_k be the relevance scores associated to the neurons at layer *l* + 1 and w_{jk} be the weight connecting neuron *j* to neuron *k*. The simple LRP rule redistributes relevance from layer *l* + 1 to layer *l* in the following way:

$$R_j = \sum_k \frac{x_j w_{jk}}{\sum_j x_j w_{jk} + \in} R_k$$

where the small stabilization term $\in$ prevents division by zero. Intuitively, this rule redistributes relevance proportionally from layer *l*+1 to each neuron in layer *l* based on two criteria, namely (i) the neuron activation x_j, *i.e.*, more activated neurons receive a larger share of relevance, and (ii) the strength of the connection w_{jk}, *i.e.*, more relevance flows through more prominent connections. Note that relevance conservation holds for $\in = 0$.

3.1.3 *Evaluation of SA and LRP*

Heatmaps produced by different explanation methods can be used to measure the quality of explanation using *perturbation analysis*. It is based on the idea that perturbing input variables important for prediction leads to a steeper prediction score decline. Input variables are sorted by relevance score, and iteratively perturbed (starting from the most relevant ones). The prediction score is tracked after every perturbation step. The average decline of the prediction score is a measure of explanation quality; a large decline indicates successful explanation.

Evaluation of SA and LRP explanations on three different problems/classifiers are described next: (i) image classification using GoogleNet, (ii) document text classification using a convolutional neural network and (iii) recognition of human actions in videos using a Fisher Linear Discriminant/SVM.

1. *Image Classification*

A deep neural network, GoogleNet, was used to classify general objects from the ILSVRC2012 dataset. Fig. 9(a) shows two images correctly classified as "volcano" and "coffee cup". The accompanying heatmaps visualize the explanations obtained with SA and LRP. The LRP heatmap of the coffee cup image shows that the model has identified the ellipsoidal shape of the cup to be a relevant feature for the category. In the volcano example, the shape of the mountain is regarded as evidence for a volcano. The SA heatmaps are much noisier than the ones computed with LRP and large values R_i are assigned to regions consisting of pure background, *e.g.*, the sky, although these pixels are not really indicative for image category "volcano". In contrast to LRP, SA does not indicate how much every pixel contributes to the prediction, but it rather measures the sensitivity of the classifier to changes in the input. Therefore, LRP produces subjectively better explanations of the model's predictions than SA. Perturbation analysis (lower part of Fig. 9(a)) shows that LRP provides better explanations than SA– due to faster prediction score decrease using LRP heatmaps than using SA heatmaps.

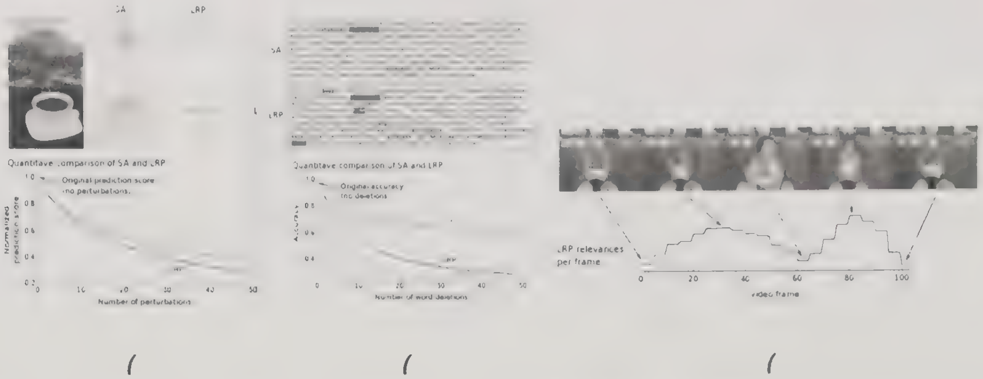


Figure 9: Explanation Quality Measures: (a) Image classification. Two images correctly classified as volcano, coffee cup by a deep learning network. Heat maps and perturbation analysis show better explanatory power of LRP than SA. (b) Text Classification: SA and LRP heatmaps identify words such as "discomfort", "body", and "sickness" as relevant ones for explaining the prediction of medicine. In contrast to SA, LRP distinguishes between positive (red) and negative (blue) relevance. (c) Human Action Recognition. Explaining prediction sit-up. The LRP heatmaps of a video which was classified as "sit-up" show increased relevance on frames in which the person is performing an upwards and downwards movement. Source: (Samek *et al*, 2018).

2. *Text Classification*

In this experiment a word-embedding based convolutional neural network was trained to classify text documents from the 20Newsgroup dataset.

Fig. 9(b) shows SA and LRP heatmaps (e.g., a relevance score R_i is assigned to every word) overlaid on top of a document, which was classified as "sci.med", *i.e.*, medical topic. Both SA and LRP indicate that words such as "sickness", "body" or "discomfort" are the basis for this classification decision. In contrast to SA LRP distinguishes between positive (red) and negative (blue) words, *i.e.*, words which support "sci.med" and words which speak for another category (e.g.,"sci.space"). Words such as "ride", "astronaut", and "shuttle" strongly speak for space, but not necessarily for medicine. With the LRP heatmap we can see that although the classifier decides for the correct "sci.med" class, there is evidence in the text which contradicts this decision. The SA method does not distinguish between positive and negative evidence. As before perturbation analysis shows that LRP provides more informative heatmaps than SA, because these heatmaps lead to a larger decrease in classification accuracy compared to SA heatmaps.

3. *Human Action Recognition*

In this experiment a Fisher Vector / SVM classifier was trained for predicting human actions from compressed videos. To reduce computation, the classifier was trained on block-wise motion vectors (not individual pixels). The evaluation was performed on the HMDB51 dataset. Fig. 9(c) shows LRP heatmaps overlaid onto five exemplar frames of a video sample. The video was correctly classified as showing the action "sit-up". The model focuses on blocks surrounding the upper body of the person as this part of the frame shows motion indicative of "sit-up", *i.e.*, upward and downward body movements. The curve at the bottom of Fig. 9(c) displays the distribution of relevance over (four consecutive) frames. The relevance scores are larger for frames in which the person is performing an upwards and downwards movement. Thus, LRP heatmaps not only visualizes the relevant locations of the action within a video frame (*i.e.*, where relevant action happens), but also identifies the most relevant time points within a video sequence (*i.e.*, when relevant action happens).

3.2 Input Features as Explanation

One approach is to attempt to explain the predictions of any machine learning classifier by having access to its input features. The goal is to provide explanations of the form "A is something because of B, C, and D." For example, "This is a bird because it has feathers, wings and a beak." Such an explanation is concise– there are not a hundred reasons. It relies on B,C,D which are also high level concepts.

Local Interpretable Model-Agnostic Explanations (LIME) is such a system (de Sousa, Vellasco, Sun, and da Silva, 2019). The use of LIME in a medical diagnostic application is shown in Fig. 10. Here the model predicts that a certain patient has the flu. The prediction is then explained by an "explainer" that highlights the symptoms that are most important to the model. The physician is thereby empowered whether to trust the model or not.

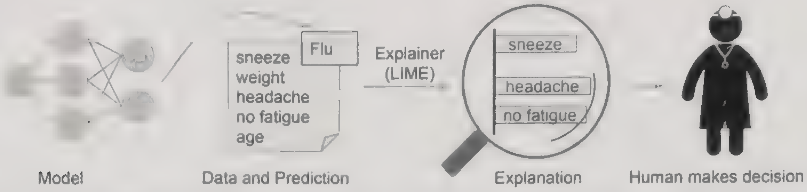


Figure 10: Input features in explaining medical diagnosis. Source: (de Sousa *et al*, 2019).

To explain a classifier that predicts whether an image contains a tree frog (Fig. 11(a)) by means of the super-pixels (a group of pixels with the same value) in the input image. The explainer generate a set of perturbed instances by turning some interpretable components "off" (making them gray) as illustrated in Fig. 11 (b). For each instance, we get the probability that a tree frog is in the image according to the model. We then learn a simple (linear) model on this data set, which is locally weighted—that is, we care more about making mistakes in perturbed instances that are more similar to the original image. In the end, we present the superpixels with highest positive weights as an explanation, graying out everything else.

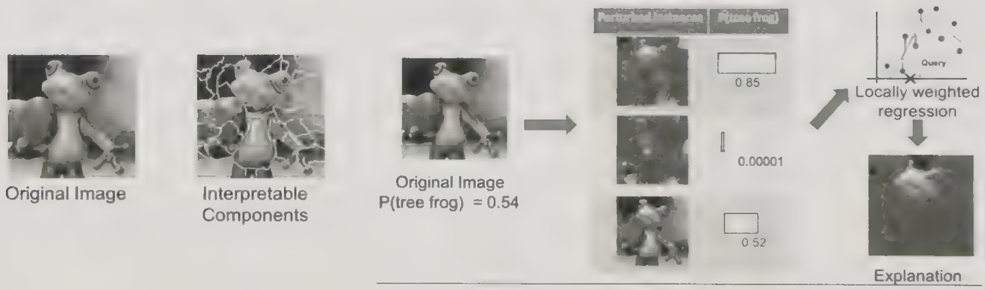


Figure 11: Input features in explaining image classification: (a) Super-pixels in input image, (b) Explaining prediction with LIME. Source: (de Sousa *et al*. 2019).

3.3 Examples as Explanation

The aim to select subset of the dataset that leads to similar conclusions as the entire dataset. The intuition is that subsets of training data that lead a model to the same (or approximately similar) inference as the model trained on all the data should be useful to understand the fitted model.

Explanation is viewed as the inverse of modeling. It is based on two fundamental observations: (i) all machine learning models are trained on data, and (ii) data is the common language of the user and the model.

Examples may not capture what philosophers and cognitive scientists call as explanation, but they harness people's proclivity to inductive inference.

It is related to machine teaching in which the teacher designs the optimal training data to drive the learning algorithm to a target model.

3.3.1 *Machine Teaching*

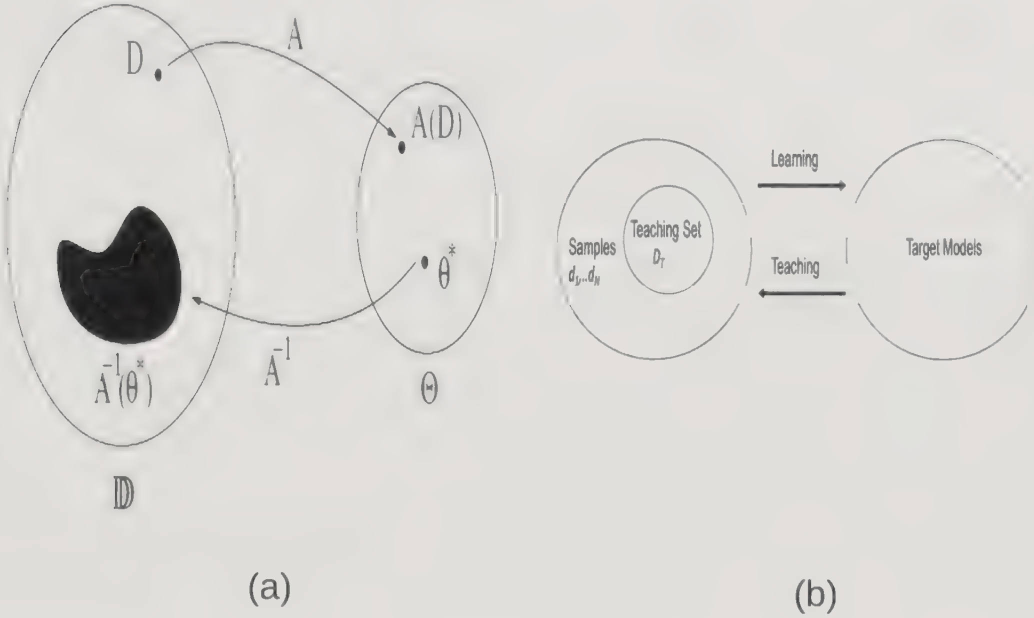


Figure 12: (a) Machine Teaching: Identifying optimal members of training set, and (b) Bayesian Teaching. Source: Yang and Shafto (2019)

Given a training dataset $D \in \mathbf{D}$, the process of machine learning returns a model $A(D) \in \Theta$. A is in general many-to-one. Conversely, given a target model $\theta^* \in \Theta$, the inverse function A^{-1} returns a set of training examples that will result in θ^*. Machine Teaching aims to identify optimal member(s) among $A^{-1}(\theta^*)$. See Fig. 12(a).

3.3.2 *Bayesian Teaching*

The teaching problem is to select a small subset of data that with high probability leads the learner to model to the correct inference. See Fig. 12(b).

This requires two kinds of inference: (i) Teacher's inference (*T*) which is done in the space of possible teaching sets, and (ii) Learner's inference (*L*) which is done in the space of possible target models. For any subset of the training data *x* the probability assigned by the model can be written as

$$P_T(x \mid \theta) = \frac{P_L(\theta \mid x)P(x)}{\int_x P_L(\theta \mid x)P_L(x)dx}$$

where θ denotes the target model, which can be an entire model or a particular substructure, such as latent features, relations, grammars, programs, or combinations of these; $P_T(x \mid \theta)$ is the probability of choosing *x* as the teaching examples for explaining target model θ, $P_L(\theta|x)$ is learner's posterior inference after receiving *x*, *P*(*x*) describes bias for certain kind of examples (*e.g.*, favoring smaller subsets), and the integral is over all partitions of the training data (*i.e.*, if the size of *x* is m and the size of the entire training corpus is *N*, there are ${}^{N}C_{m}$ partitions).

A Bayesian teacher:

With training data $D = \{d_1, d_2, , d_N\}$ and teaching set size $n < N$ teaches a target model θ^* by sampling a teaching set $D_T \subseteq D$ from $D = \{D \mid D \in P(D) \wedge | D \leq n\}$ according to

$$p(D_T \mid \theta^*) = \frac{p(D_T)p_L(\theta^* \mid D_T)}{p(\theta^*)} = \frac{p(D_T)p_L(\theta^* \mid D_T)}{\sum_{D \in D} p(D)p_L(\theta^* \mid D)}$$

where *P(D)* is the power set of *D* and *D* is the space of teaching sets. $p_L(\theta^* \mid D)$ is the probability the learner will infer the target model θ^* given a particular teaching set *D* (*i.e.* the learner's posterior probability given that teaching set), and *p(D)* is the teacher's prior probability on the same teaching set *D*. Priors assign higher probabilities to smaller teaching sets.

Teaching Inference for model explanation is as follows. Pick a teaching set size (*e.g.*, 2) to constrain the search space. Perform teaching inference for the category to be understood. Rank the teaching sets based on the teaching probabilities. See Fig. 13.

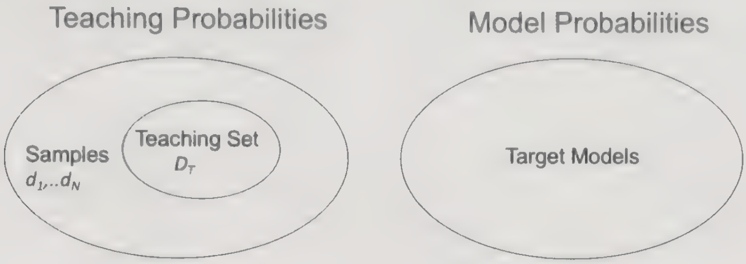


Figure 13: Teaching Inference for model explanation. Source: (Yang and Shafto, 2019).

Consider the explanation of a new model at a conference. Authors show examples classified correctly and incorrectly. They mention that humans would find some to be hard. Rather than cherry-pick, we care about those classified correctly with high confidence, those classified correctly with low confidence, those classified incorrectly with low confidence, those classified incorrectly with high confidence. But some methods don't offer certainty estimates. Bayesian teaching offers finding/ranking such examples as seen below.

The five best teaching sets using *ground truth labels* are shown in Fig. 14 (a). Each pair of images represents a teaching set; pairs are sorted by teaching probabilities in descending order, from left to right (leftmost is best). The five worst teaching sets using ground truth labels are in Fig. 14 (b). Each pair of images represents a teaching set; pairs sorted by teaching probabilities in ascending order, from left to right, (leftmost set is worst). The five best teaching sets using *model predictions as labels* are shown in Fig. 15(a). The five worst teaching sets using model predictions as labels are in are shown in Fig. 15(b).

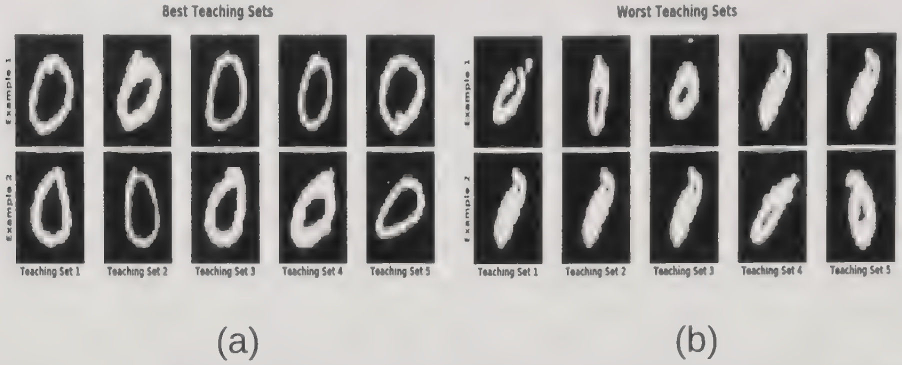


Figure 14: Using ground truth as labels: (a) Best teaching sets, and (b) Worst teaching sets. Source: (Vong, Sojitra, Reyes, Yang, and Shafto, 2018).

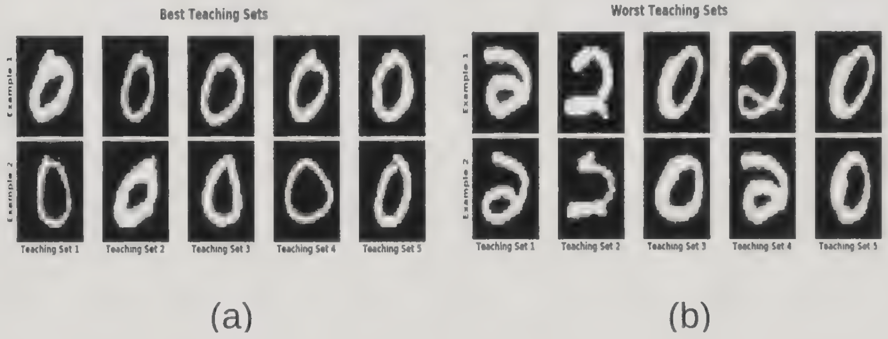


Figure 15: Using model predictions as labels: (a) Best teaching sets, and (b) Worst teaching sets. Source: (Vong *et al,* 2018).

The model for Category 0 is explained by: (i) those classified correctly with high confidence; (ii) those classified correctly with low confidence; (iii) those classified incorrectly with low confidence; and (iv) those classified incorrectly with high confidence.

In summary, Bayesian teaching leverages the common understanding of model behavior—the data—to explain opaque models through the examples from the original data that are most representative of the inference. In doing so, it integrates learning and explanation by taking the learning model as input into the explanation process, and outputs an explanation in terms of the examples from the original data.

4 Ante-Hoc XAI

While post-hoc explanation techniques allow models to be trained normally, with explainability being incorporated as an afterthought, ante-hoc techniques entail making explainability into a model from the beginning. The goal of tranisitioning from a conventional AI system to an ante-hoc AI system, in the context of image recognition, is illustrated in Fig. 16.

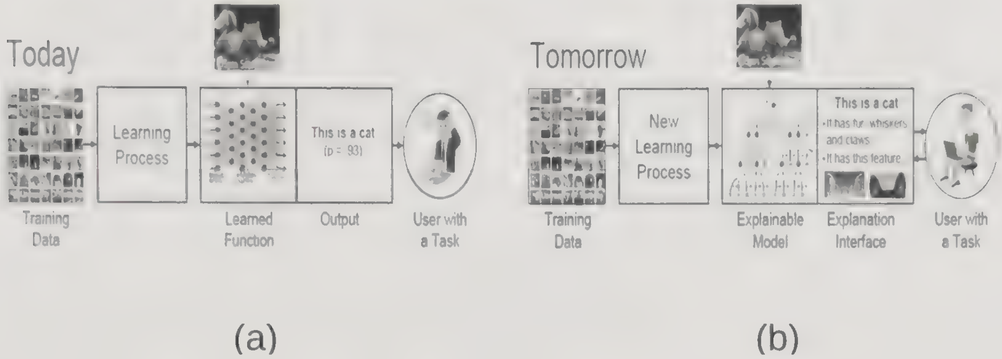


Figure 16: Goal of Ante-hoc XAI. Source: (Turek, 2018).

The goal of ante-hoc AI is to produce more explainable models, while maintaining a high level of performance, say prediction accuracy. The goal is to enable human users to understand, trust and manage emerging AI partners. Do we want a complex black box model such as an RNN or a less accurate traditional model with better interpretation, say logistic regression, *i.e.*, a 90% accurate model we understand versus a 99% accurate model we don't. The role of performance in ante-hoc AI is illustrated in Fig. 17.

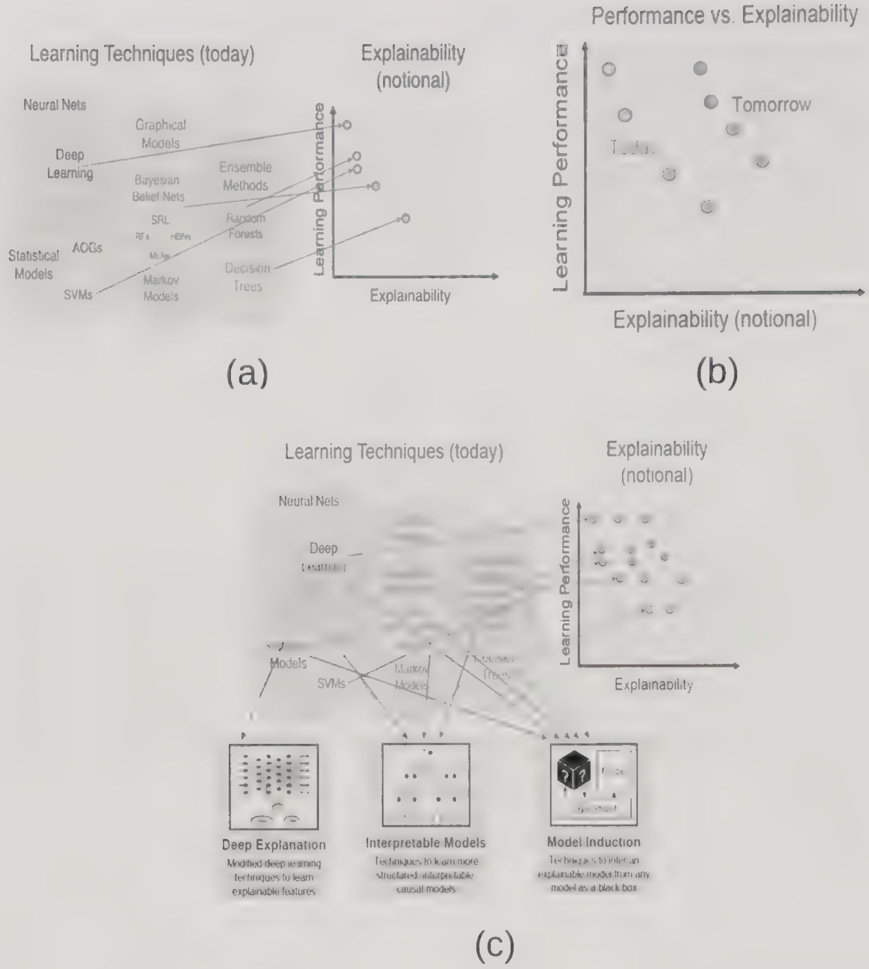


Figure 17: Performance of Explainable AI: (a) today, (b) tomorrow and (c) tomorrow's methods. Source: (Marselis, 2019).

Some active research efforts in ante-hoc AI are:

- Explanation from Representation. Techniques to identify the most salient input features used in a decision, *e.g.*, it is a cat because it has whiskers and fur. They include techniques to select the training examples most influential in a decision. The explanation can also be based on embedded or computed features, *e.g.*, network dissection techniques to identify meaningful features inside the layers of a deep net. Intertwined are deep learning techniques to generate explanations.

- Probabilistic Explanation. These methods leverage inference methods from probabilistic graphical models. They are a natural approach to use with neuro-symbolic methods.

In the following two subsections we describe efforts based on each of these two approaches.

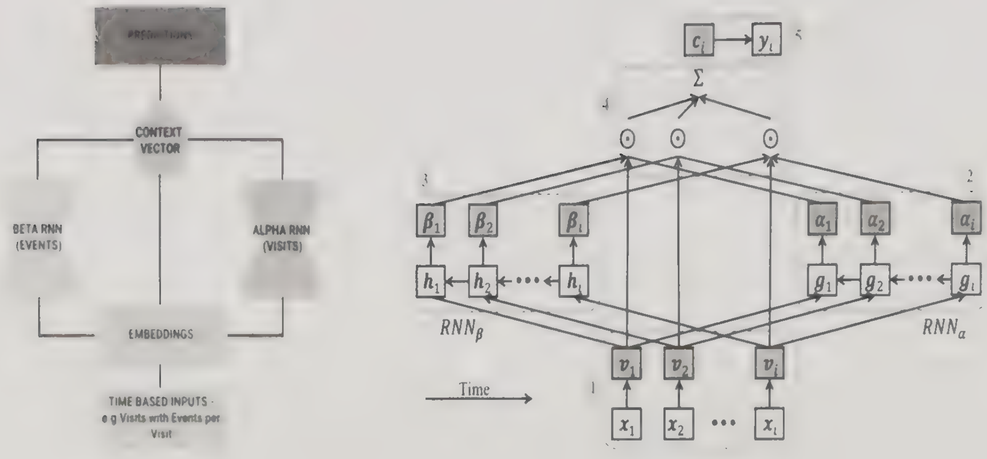
4.1 Explanation from Representation

4.1.1 *Reverse time Attention Model*

The Reverse time Attention Model (RETAIN) mimics a physician in providing explanations (Choi et al, 2016). The goal is to help physicians understand the AI software's predictions. RETAIN uses an electronic health record (EHR) in reverse time order. It calculates the contribution of variables (medical codes) to diagnostic prediction using RNNs. See Fig. 18.

Patient hospital visit data is sent to two recurrent neural networks (RNNs) both of which have an attention mechanism. The concept of attention in RETAIN is analogous to attention in machine translation. Given a sentence of length S in the source, first generate $h_1,\ldots,h_S$ to represent input words. Then to find the j^{th} target word, generate attention α_i for $i=1,\ldots,S$ for each word in the source sentence. Compute context $c_j = \sum_i \alpha_{ij} h_i$ and use it to predict the j^{th} target word i. Attention allows focus on specific words in the given sentence when generating each word in the target.

The attention mechanism in RETAIN helps explain which part the neural network was focusing on and which features helped influence its choice.



(a) (b)

Figure 18: Reverse Time Attention Model: (a) overview, and (b) detail of RNNs and context vector. Source: (Choi *et al*, 2016).

4.1.2 *Explanations from Embeddings*

The essence of deep learning is that of learning representations, or embeddings, that are useful to easily perform the final computation task, of, say classification or regression.

A proposed approach to harness what has been learnt by a deep network is to construct an explanation module by embedding a high-dimensional deep network layer nonlinearly into a low-dimensional explanation space. In this process a goal is to retain faithfulness, so that the original deep learning predictions can be constructed from the few concepts extracted by the explanation module.

The explanation module is a dimensionality reduction mechanism so that the original deep learning prediction $\hat{y}$ can be reproduced from this low-dimensional space. It can be attached to any layer in the prediction deep network. The network output can be faithfully recovered from this low-dimensional explanation space. A sparse Reconstruction Autoencoder is used as an explanation module (Qi, Khorram, and Li, 2018). See Fig. 19(a). An explanation generated by this model is shown in Fig. 19(b).

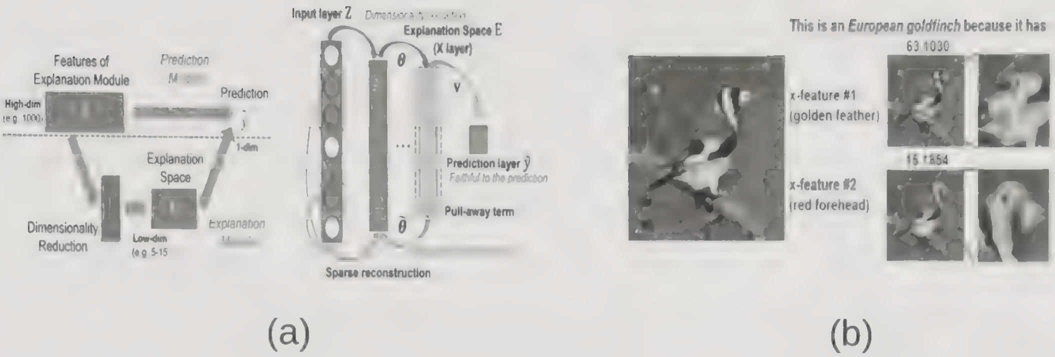


Figure 19: Generating visual explanations: (a) a sparse reconstruction autoencoder used to generate explanations, and (b) an explanation generated. Source: (Qi *et al*, 2018).

Caption-guided Image explanation

Deep image captioning systems learn to translate visual input into languages: potential map between visual concepts and words. Despite good captioning performance, they are hard to understand "black boxes." A solution proposed is caption guided visual saliency: a top-down neural saliency map. See Fig. 20.

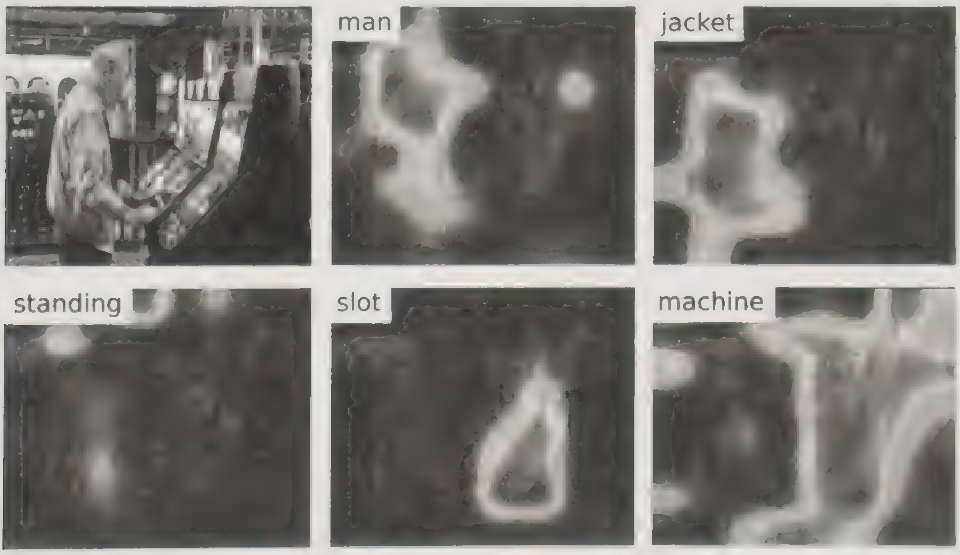


Figure 20: Caption guided visual saliency. Source: (Ramanishka, Das, Zhang, and Saenko, 2017)

4.2 Probabilistic Explanations

4.2.1 *Bayesian Deep Learning*

Bayesian deep learning (BDL) enables one to gauge how uncertain a neural network is about its predictions. These deep architectures can model complex tasks by leveraging the hierarchical representation power of deep learning, while also being able to infer complex multi-modal posterior distributions. BDL models typically form uncertainty estimates by either placing distributions over model weights, or by learning a direct mapping to probabilistic outputs. By knowing the weight distributions of various predictions and classes, we can tell a lot about what feature led to what decisions and the relative importance of it.

4.2.2 *Graphical Model Inference*

Neuro-symbolic models aim to use both neural mechanisms to infer symbolic entities, but also aim to incorporate symbolic reasoning mechanisms to answer queries of interest. For instance, in an image of a horse, the neural mechanism infers that we have a horse with high probability. Before performing classification it identifies its features such as its legs, tail, mane, etc. A simple linear classifier performs the final classification. Suppose one of the horse's legs is occluded. Even though the neural model can infer the presence of a horse, it would require a symbolic mechanism to infer the presence and location of the invisible fourth leg, and provide an appropriate explanation (Fig. 21).

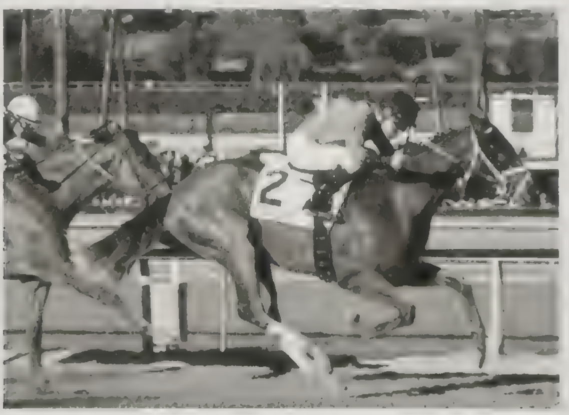


Figure 21: Image of two horses with one having its legs cropped out. While a neural network would recognize both horses, based on observed features, a symbolic reasoning mechanism would infer the presence and location of the cropped-out legs. Image source: Benoit photo.

Symbolic reasoning and inference is done efficiently using graphical models such as Bayesian networks and Markov networks. In fact the inference of the most likely probability of a set of variables in a graphical model given evidence is referred to as maximum a posteriori (MAP) probability inference equivalently referred to as the most probable explanation (MPE) of the evidence (Koller and Friedman, 2009).

Probabilistic graphical models represent joint probability distributions over a set of variables χ. They are used to answer queries of interest, given evidence variables whose known value is *e*, *i.e.* $E = e$. Assume that a query to the system yields a value y to the target variables *Y*. We assume that the explanatory variables are the rest of the variables $W \subseteq \chi - Y - E$.

Assume $Y = y$ is the response to the query variable. The conditional probability of explanation is $P(W \mid E = e,\ Y = y)$. The task of MPE can be expressed as follows.

$$\arg \max_{w \in W} P(W \mid E = e, Y = y)$$

which can be computed using

$$P(w \mid E = e, Y = y) = \frac{P(w, e, y)}{\sum_e \sum_y P(w, e, y)}$$

Since the denominator has an exponential number of terms in the number of variables, the computation is NP-hard. The situation can be alleviated by using methods from probabilistic graphical model domain. They include exact inference algorithms for MPE such as *variable elimination* and *clique trees*. Also, approximate algorithms based on either optimization or sampling (particle-based).

5 Concluding Remarks

Among the large number of research areas spawned by the growth of applications of artificial intelligence, one that has attracted general public interest is that of trustworthiness. Among key attributes of trustworthiness is explainability. An attribute that can be useful not only to the user, *e.g.*, to understand whether there is bias, but also to the designer, *e.g.*, to improve design.

Explanations can be made in terms of identifying salient features in the input. They can also be in terms of computed features, known as embeddings in deep learning systems. Explanations of AI systems can be made in terms of identifying archetypes in the training data.

If explanations are generated from existing AI systems, they are called post-hoc systems. AI systems configured from the start to generate explanations are known as ante-hoc systems.

Post-hoc explanations can be generated by perturbing input variables in variable heat maps so as to determine as to which variables have the most effect in changing the generated decision. They can also be generated from deep learning models using attention mechanisms that identify the context for each decision.

Ante-hoc explanations have been obtained by mapping outputs of layers to lower-dimensional spaces. They can also be generated using neuro-symbolic methods, where a neural network generates values of high level variables and a symbolic reasoning system generates an explanation in those terms.

Probabilistic graphical models offer computationally feasible algorithms to generate probabilistic explanations in terms of given evidence and the observed target value.

6 References

[1] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 3504–3512. Curran, 2016.

[2] I.P. de Sousa, M. Vellasco, J. Sun, and E.C. da Silva. Local interpretable model-agnostic explanations for classification of lymph node metastates. Sensors, 19(3), 2019. https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/.

[3] National Science Foundation. National Artificial Intelligence (AI) Research Institutes, 2019.

[4] I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. MIT Press, 2016.

[5] L.A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In In Proceedings European Conference on Computer Vision, 2016.

[6] D. Koller and N. Friedman. Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009.

[7] J. Mao, C. Gan, P. Kohli, J.B. Tenenbaum, and J. Wu. The neurosymbolic concept learner. In In Proceedings of ICLR, 2019.

[8] R. Marselis. Make your Artificial Intelligence more trust- worthy with eXplainable AI.

[9] K. Pei, Y. Cao, J. Yang, and S. Jana. Deepxplore automated whitebox testing of deep learning systems. In Proceedings 26th Symposium on Operating Systems Principles, 2017.

[10] Z. Qi, S. Khorram, and F. Li. Embedding Deep Networks into Visual Explanations, 2018.

[11] V. Ramanishka, A. Das, J. Zhang, and K. Saenko. Top-down Visual Saliency Guided by Captions, 2017. https://arxiv.org/abs/1612.07360.

[12] W. Samek, T. Wiegand, and K.R. Muller. Explainable artificial intelligence: Understanding,visualizing and interpreting deep learning models. ITU Journal: ICT Discoveries, pages 39–48, 2018.

[13] M. Turek. Explainable Artificial Intelligence (XAI), 2018. https://www.darpa.mil/program/explainable-artificial-intelligence.

[14] W.K. Vong, R.B. Sojitra, A. Reyes, S. Yang, and P. Shafto. Bayesian Teaching of Image Categories. scottchenghsinyang.com/paper/Vong-2018.pdf.

[15] S.Y. Yang and P. Shafto. Explainable Artificial Intelligence via BayesianTeaching. http://shaftolab.com/assets/papers/yangShafto_NIPS_2017_machine_teaching.pdf.

BIO

Sargur Srihari is SUNY Distinguished Professor of Computer Science and Engineering at the University at Buffalo, where he teaches and conducts research in Machine Learning and Artificial Intelligence. Srihari led a team that developed the world's first automated system for reading handwritten postal addresses. His honors include: Fellow of the Institute of Electronics and Telecommunications Engineers (IETE, India), Fellow of the IEEE, Fellow of the International Association for Pattern Recognition and distinguished alumnus of the Ohio State University College of Engineering.

Commonsense and Explanation: Synergy and Challenges in the Era of Deep Learning Systems

Gary Berg-Cross

Ontology Forum Board Member

Abstract

This article builds on some of the research ideas discussed in the commonsense reasoning and knowledge track as part of the Ontology Summit 2019 on explanations. As discussed there, research on intelligent systems has long emphasized the benefits of providing explanations for system reasoning, although approaches to an explanation function have evolved over time. While system-provided explanations like common-sense knowledge (CSK) and associated reasoning (CSKR) each go back to the early days of artificial intelligence (AI) systems, they became somewhat independent research areas for much of their later history. This was in part because explanations in early AI efforts were technical in nature centering on how faithfully a system describes the reasoning and heuristic steps employed. Another factor was the difficulty of building adequate bases of CSK for reasoning. Although early AI notionally recognized that as part of intelligent systems explanations should make commonly understood sense, this was not a sustained priority in later work. Instead CSK research engaged more on issues of adequate knowledge representation, how to acquire a base of CSK and the diversity of ontologies needed to support CSK. While these are not finished research areas, they now provide useful guidance to support a current interest in the role of CSK explanations motivated by new challenges and opportunities. These include the rapidly expanding space of heterogeneous and richly interconnected data along with diverse sub-symbolic (deep learning) intelligent system applications. New AI approaches include useful, but only partially understood results, from machine learning (ML) and deep neural net (DNN) approaches. The complexity of these approaches, which includes use of patchy and inconsistent information available online, prompts a renewed desire to have systems explain their decisions and processing in deep, flexible, defendable and understandable ways. Recent work has promoted the development of AI systems using ML-based models with a range of explanatory capabilities for generated decisions. Common sense concepts now play a role in providing better performance and a range of more easily understood explanations for end users.

Taken as a whole, the cumulative lesson of decades of research is that fluid explanations, responsive to changing circumstances require knowledge about the world and that explanations are intimately connected to both common-sense reasoning and background knowledge such as captured in formal ontologies, but also informally understood in text (Davis and Marcus, 2015). The combination of information and its context extracted from a range of sources and organized and represented formally provides a base, not only for intelligent system performance, but also for background knowledge needed for flexible and deep explanations. In practice there seem to be many views of satisfactory explanations but that CSK and reasoning plays varying, but useful roles in each of these.

Among the remaining challenges are those of developing an adequate base of CSK, an adequate approach to situational and contextual understanding, how to use deep learning in dynamic situations, the need to keep humans in the loop and the need for a common enhanced ontology engineering practice addressing both explanation and CSK.

Introduction

The 2019 Ontology Summit on Explanation (Ontology Summit, 2019) provided an opportunity to look at various approaches of intelligent/smart systems [i]from a number of perspectives including that of commonsense knowledge (CSK) and associated reasoning (CSKR). Commonsense reasoning and knowledge was prominently featured as an early part of Artificial Intelligence (AI) conceptualization, and it was assumed to be important in the development and enhancement of human-like, intelligent systems explanations, which also had a defined role in early AI. Both continue to be considered important parts of intelligent systems and this is not surprising when we consider the centrality of an ability to explain reasoning and what they know by a system whose claim to fame is intelligence itself. Over the past half century of work on intelligent systems, a variety of approaches to explanation have been engineered and deployed and when carefully designed proven useful. On the whole CSKR's role in explanations has been more indirect than direct. It has often been used to provide a perspective on explanatory short comings. However, new ML techniques that construct and represent knowledge using non/sub-symbolic models layer additional requirements for understandable explanation. This in turn provides and opportunities for CSKR to aid in such explanations (Chakraborty, *et al.* 2017).

In the sections that follow I discuss some of the historical relations of explanation and CSKR followed by some of the experience over time of crafting both CSKR and good explanations. A useful way to illustrate the current status of this work is to overview how some explanation applications are built and employed in representative areas. Following this I overview some of the issues and some of the challenges introduced by a consideration of applying CSKR to contemporary AI and ML systems and the recent efforts in the new field of eXplanatory AI (XAI). We conclude with a summary of some preliminary findings, identification of remaining issues and opportunities that might promote and guide future work.

Some Background on the AI Connections of Commonsense and Explanation

In this section I review some of the major developments along the AI path to intelligent systems and why CSKR seems like an important ingredient in the development of intelligent explanation. Note, that this review is not comprehensive, but represents a survey giving the flavor of methods and results that are pertinent to the evolution of explanations and CSKR.

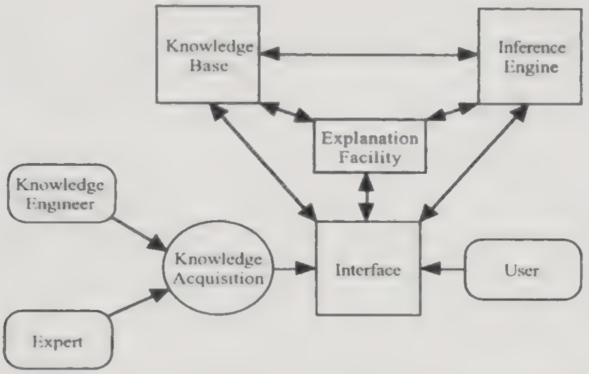
Simply put, fifty years of experience teaches us that only an intelligent system that justifies its actions in terms which make sense so they are readily understandable to the user will be trusted (Cohen *et al*, 2017) . Early AI work showed that rudimentary attempts at explanation provided useful to system engineers and a modicum of user satisfaction if not trusted (Langlotz and Shortliffe, 1984). As a result improvements in explanation have remained a necessary next step in intelligent system evolution for a long time. Interestingly, one sees in the original Turing Test the need for CSKR and explanations each as part of communication to pass the test. These are, of course, common human abilities to live in an ordinary world (Ortiz, 2016). Some examples of CSKR needed for passing a Turing Test or just living in society are illustrative of the range involved and might include the following type of reasoning:

- Taxonomic: Cats are mammals.
- Causality: Eating a peach makes you less hungry.
- Goals: I don't want to get hot so let's find shade.
- Spatial: You often find a microwave in the kitchen.
- Functional: You can sit on a chair if tired.
- Planning: To check today's weather look on a weather application.
- Linguistic: The word "won't" is the same as "would not".
- Semantic: Cat and feline have a similar meaning.

Many cognitive abilities that are developed it seems simply in the first years of life provide the commonsense knowledge and reasoning to handle the above list and problems like conservation of objects - if I put my toys in the drawer, they will still be there tomorrow. It has proven much harder to get such an adequate base of knowledge and associated reasoning into computational systems. Early on in this process two ways seem possible to populate such knowledge for an intelligent system. One is by handcrafting in a mass of commonsense knowledge, while another is by letting a system learn from training experience with things like object conservation over time or place. One may also consider some combination of the two, say building in some knowledge and using that to learn more, or letting it learn and correcting errors by adding hard to learn knowledge or by dialog with a user.

Indeed an early AI goal was to endow systems with natural language (NL) understanding and text production, which it worth noting could be used for explanations. It is easy to see that a system with both CSKR and NL facilities would be able to provide smart advice as well as explanations of this advice. We see both in the early conceptualization of a smart advice taker system from McCarthy's work making causal knowledge available for: "a fairly wide class of immediate logical consequences of anything it is told and its previous knowledge." McCarthy (1960) further noted that this useful property, if designed well, would be expected to have much in common with what makes us describe certain humans as "having commonsense." John McCarthy believed so and argued that a major long-term goal of AI should include endowing computers with standard commonsense reasoning capabilities.[ii]

While there is a long history showing the relevance of commonsense knowledge and reasoning to explanation in actual practice, going back to the 70s and 80s, AI systems, aka "expert systems", were not as the founders envisioned. They were less knowledgeable and brittle, based on explicit models of domains implemented using handcrafted production rules encoding useful information about special topics such as diseases. In part because of handcrafting of knowledge rather than the engineering of knowledge systems rule, knowledge was fragmented and opaque and would break down revealing obvious errors. Part of this was due in part to a lack of the robustness available from human-like commonsense which was hard to handcraft or engineer into applications' supporting knowledge bases. Following an easier development path 1970s era expert systems came, as shown in Figure 1, with a very simple, technical, but not commonsense rich idea, of what was called an "explanation facility." The early implementations used a proof trace of rule firings which provided a purely technical explanation. It did not include what we call justifications for its explanations. Such proofs founded on "Automated Theorem Provers" (Melis, 1998) could provide a map from inputs to outputs and served the needs of system engineers to understand system performance more than providing an explanation to a user[iii].



Example of an Early AI System Architecture
From Medsker, Larry R. Hybrid intelligent systems.
Springer Science & Business Media, 2012.

Figure 1. Simple View of an Early Expert System

But case specific and mathematical based proof planning are not as robust or as reliable as they first seemed to AI developers. This was due to the commonly understood fact that situations being reasoned over were often not adequately represented. Thus, situations and the explanations about them lose some intuitive meanings expected by users (Bundy, 2002). Another problem is that rules in a knowledge base (KB) can change over time and early efforts did not include meta-knowledge to explain why they change. To make sense changes often need explanations.

Along with brittleness and limited utility of traces, part of the weaknesses of rule-based explanatory reasoning, was exposed by Clancey (Clancey, 1983). He found that the AI system called Mycin's had individual rules that play vastly different roles, have different kinds of justifications, and are constructed using different rationales for the ordering and choice of premise clauses in the rules. Since in this rule knowledge isn't made explicit, it can't be used as part of explanations. And there are structural and strategic concepts which lie outside of early AI system rule representations. It was soon realized that these can only be supported by appealing to some deeper and contextual level of (background) knowledge. But commonsense context was seldom "explicitly stated" and thus difficult to engineer.

In searching for solutions the next generation of AI developers used more structured and formal KBs such as frames or semantic net-like ontologies to capture and formalize a fuller range of necessary knowledge. At this time the role of causal-based explanations also helped design more knowledgeable and integrated rather than *ad hoc* expert systems, based on the idea that a system's knowledge should be integrated with performance and adequate to explain its reasoning (Swartout and Smoliam, 1989). Taken together this made the argument that something like ontologies are needed to make explicit structural, strategic, and support-type knowledge. One result was development of large KBs such as in the Cyc project (Lenat and Guha, 1989), a 35-year effort to codify common sense into an integrated, logic-based system. Efforts like Cyc which started up in the 80s represented an effort to avoid problems like system brittleness by providing a degree of common-sense and modular knowledge (Lenat, *et al* 1985). Cyc can provide a response to queries such as: "Can the Earth run a marathon?" In terms of a commonsense explanation we have a "no" because of the knowledge that the Earth is not animate and the role capability needed to run a marathon is detailed by the knowledge in a sports module. Indeed the need for a formal mechanism for specifying a commonsense context had become recognized, and some approach to it, such as Cyc's microtheories arose[iv]. In the 80s Cyc-type knowledge was also seen as important to what was called associate systems. This advance argued that "systems should not only handle tasks automatically, but also actively anticipate the need to perform them....agents are actively trying to classify the user's activities, predict useful sub-tasks and expected future tasks, and, proactively, perform those tasks or at least the sub-tasks that can be performed automatically" (Panton, 2006). All of these abilities were, of course, conceptually useful for explanation, so advances in CSKR, like a Cyc micro-theory could serve a dual role. Much ontological work has followed the spirit of this idea if not the exact program outlined to build large KB such as the Cyc project.

But subsequently, except for a few systems they were rarely applied as part of mainstream systems although the need was often noted (Minsky, 2000). Although some efforts, such as Crowdsourcing common sense training data in Open Mind (Singh, 2002a) are notable, the effort to engineer sufficient CSK for reasoning as well as reasonable explanations has proven difficult. While there are some success as an aid to NLP, where hybrid approaches out perform an NLP tool like BERT (Havasi, 2019), the scale of the problem has been discouraging; for people seem to need a tremendous amount of knowledge of a very diverse variety to understand even the simplest children's story (Singh, 2002b). Research retreated from an ambitious broad CSKR aim and instead pursued special domain knowledge and reasoning that could deal with a more focused class of problem. But these lacked generalization and thus did poorly at almost everything else (Minsky, Marvin L., Push Singh, and Aaron Sloman, 2004).

Despite direct approaches to explanation and problems of formalizing background knowledge, work since the 1990s has included other forms, styles, or meanings of explanation that seemed easier. Because proof isn't always useful and deep background knowledge is hard to formalize another form of documentation, and thus a style of explanation has often been used that involves the provenance or source of some fact or statement (McGuinness, 2003, Moreau, 2010, Darlington, 2013). This arises often when we want an explanation to make clear what the documented source of data is[v].

In contrast AI explanation work in the 90s and early 2000s focused on simpler techniques to make explanations acceptable to novice users rather than using large KBs of CSK which were expensive, time consuming, and hard to build with the tools and limited expertise available. Modest use was made of cognitive learning theory and associated technology[vi] which suggested the need for explanation justification using explicit knowledge of things like conceptual terminology, domain facts, and causal relations to enhance the ability of novice user's understanding (Darlington, 2013). What was more desired was explanations that also aided engineers in modifying systems (*e.g.* knowledge debugging as part of KB development).

CSKR and Explanation in the new era of ML

As noted earlier, it can be costly to acquire an adequate base of CSKR for its own sake as well as leverage it for explanations. And, when acquired, since there are a variety of ways to represent CSKR, from symbolic forms of rules to semantic nets, and logic, the knowledge content becomes heterogeneous and siloed making them difficult to integrate and structure for explanations.

This makes it attractive to consider lighter methods for acquiring knowledge like opportunistic extraction processes from text, including online text and linked data, using AI, ML, or NLP tools. Rapidly advancing ML capabilities have raised the hope of capturing knowledge including masses of CSK in more automated ways that are less resource intensive. There has now been a decade of work to acquire and represent domain knowledge, even some commonsense-like knowledge, using automated extraction and ML processes that acquire models learned from training data. A remaining problem with early work that is still somewhat with us is that a large store of training data is needed because the model must learn anew from scratch each time it learns anything. And this isn't how people work.[vii]

One prominent, illustrative attempt to tackle this problem is the Never-Ending Language Learner (NELL) system which uses a coupled semi-supervised training approach (Mitchell *et al*, 2018). Central to the NELL effort is the idea that we will never truly understand machine or human learning until we can build computer programs that share some similarity with the way humans learn. This does promise the possibility of acquiring a useful set of CSKR along the way. In particular such systems, as discussed by (Mitchell *et al*, 2018), are like people in that with years of diverse, mostly self-supervised experience, they can learn many different types of everyday knowledge or functions and thus information from many contexts. This happens in a staged bootstrapping fashion, where previously learned knowledge in one context enables learning further types of knowledge. It is easy to elaborate on cognitive processes for informed ML (Von Rueden et al, 2019) using ideas such as self-reflection on existing knowledge and the ability to formulate new representations and new learning tasks that enable the learner to avoid stagnation and performance plateaus.

As reported in Michell *et al* (2018) NELL has been learning to read the web 24 hours/day since 2010, and at that time had acquired a knowledge base with over 80 million confidence weighted beliefs (*e.g.*, servedWith(tea, biscuits).90 confidence). NELL has also learned millions of features and parameters that enable it to read these beliefs from the web. Additionally, it has learned to reason over these beliefs to infer (we might say using CSKR) new beliefs, and is able to extend its ontology by synthesizing new relational predicates. NELL learns to acquire knowledge in a variety of ways. It learns free-form text patterns for extracting this knowledge from sentences on a large scale corpus of web sites and it learns probabilistic rules that enable it to infer new instances of relations from other relation instances that it has already learned[viii]. As an example, NELL might learn a number of facts from a sentence defining "icefield", such as:

"a mass of glacier ice; similar to an ice cap, and usually smaller and lacking a dome-like shape; somewhat controlled by terrain."

In the context of this sentence and this new "background knowledge" extracted it might then extract supporting facts/particulars from following sentences:

"Kalstenius Icefield, located on Ellesmere Island, Canada, shows vast stretches of ice. The icefield produces multiple outlet glaciers that flow into a larger valley glacier."

Also of importance is that not only the textual situation is used to inter-relate extracted facts, but the physical location (*e.g*., Ellesmere Island) and any temporal situations expressed in these statements is used as context.[ix] NELL remains an example of how NLP and ML approaches can be used to build CSK and domain knowledge, but source context as well as ontology context needs to be taken into account to move forward. But NELL while it has extensive knowledge, it has relatively shallow semantic representations and thus suffers from ambiguities and inconsistencies (Gunning, 2018). And compared to handcrafted information such parts of extracted information are inconsistent with other parts and much noisier. Further, it is challenging to capture relevant situational context which include potentially important relations to other concepts - much of what is needed may be implicit and inferred and is currently only available in unstructured and un-annotated forms such as free text. And often training inputs to the model are highly engineered features that are complex or difficult to understand, meaning the resulting model learned will be hard to decompose for understanding use as input to explanation.

But progress on this problem comes from advanced ML applications where prior knowledge (background knowledge) may be used to judge the relevant facts of an extract, which makes this a bit of a bootstrapping situation.

Despite the remaining problems it seems reasonable that the role of existing and emerging Semantic Web technologies and associated ontologies is central to making CSKR population viable and that some extraction processes using a core of CSKR may be a useful way of proceeding.

CSKR helps Understanding and Thus Performance as well as Explanation in Contemporary ML Applications

As we have seen, the context that is important for discussing contemporary approaches to CSKR and explanations is that AI systems increasingly use advanced techniques such as deep learning (DL). These may in turn require additional techniques to make them more understandable to humans and system designers as well as trusted. For a different reason the current excited emphasis on explanation grows in part out of a feature failure of Deep Learning (DL) solutions - without additional effort they are opaque, at least in the sense that the models learned are not transparent to users or engineers. Despite this, contemporary deep neural networks (DNNs) have seemingly achieved near-human accuracy levels in various types of classification and prediction tasks including image and object recognition, text, speech, video data and behavior monitoring. These are all considered "low-level" tasks and advanced operations like planning or focused attention are not involved. Like simple rule-based explanations before them, raw DL systems do not natively handle desired aspects of explanations. Post-hoc explainability may be added to make them seem responsive. More recently, researchers, such as part of DARPA's XAI program, as described by (Srihari, 2020) in this Issue, aim to create a suite of rich ML techniques that:

- produce more explainable models while maintaining a high level of learning performance and,
- enable humans to understand, appropriately trust, and effectively manage the emerging generation of AI "associates" that can be used in "high-level" domains such as healthcare, criminal justice system, and finance (Goodfellow 2016).

A notional architecture of a modern, hybrid intelligent system is shown in Figure 2. Here knowledge and reasoning are divided into several types which produce not only better problem solving abilities but explanations interpretable to a range of audience types. In order to achieve this a range of knowledge sources is involved as well as ML applications to further enrich the acquisition process.

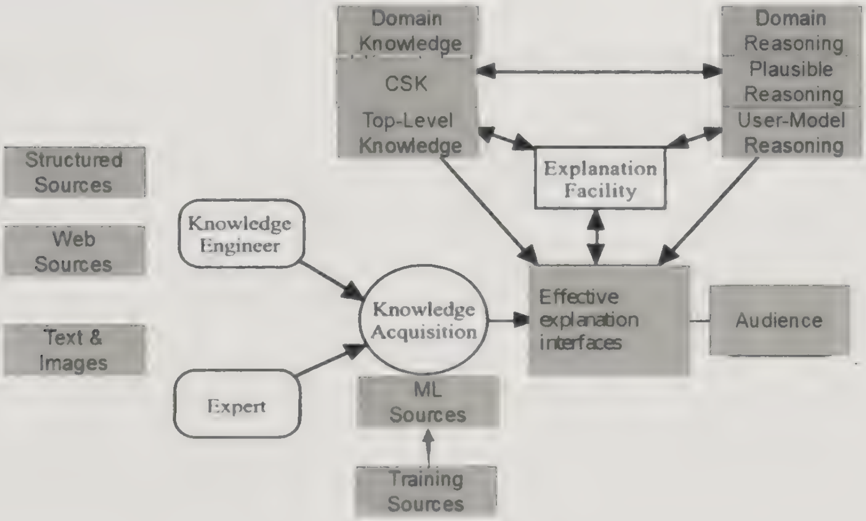


Figure 2: Architecture of a Hybrid Intelligent System

As an example, until recently the networks developed by ML for even simple vision detection approaches were treated mostly as black-box function approximators, in which a given input is mapped to some classification output such as the task of labeling images or translating text, as discussed in tracks of the 2019 Ontology Summit (Baclawski, 2018). So while ML and DL applications are now in wide use for common tasks such as advanced navigation with some sort of explanations to users, they are not naturally conducive to the generation of explanation structures. Because of complexity model simplification, say creating a decision tree, and/or feature relevance techniques which gauges the influence, relevance, or sensitivity each feature has in the prediction output by the model to be explained.

Supplications are often needed as a basis for explanations (Arrieta *et al*, 2020). Thus while non-technical, but valid and commonsense fashion are increasingly desired, they do not come without additional effort. Yet as Gunning summarizes:

> Machine reasoning is narrow and highly specialized. Developers must carefully train or program systems for every situation. General commonsense reasoning remains elusive. The absence of common sense prevents intelligent systems from understanding their world, behaving reasonably in unforeseen situations, communicating naturally with people, and learning from new experiences. Its absence is perhaps the most significant barrier between the narrowly focused AI applications we have today and the more general, human-like AI systems we would like to build in the future. (Gunning, 2018)

In a sense this is a return to an early desire to have smart applications knowledgeable about common phenomena and coincidentally ones capable of providing satisfying, interpretable explanations, but now positioned to take advantage of AI advances using DL. The path is necessary even though we still have not solved all the challenges of CSKR. Considering the range of application anticipated the goal of a reasonably competent CSKR system should include the ability to reason about explanations ("that makes sense") taking into account things like predictions, generalization, metaphors and abstractions, examples, as well as the goodness of plans, and diagnosis.

There is an obvious trust benefit if semi- or fully-automatic explanations can be provided as part of decision support systems. This seems like a natural extension of some long used and understood techniques such as logical proofs. Benefits can easily be seen if rich and deep deductions could be supported in areas regarding policies and legal issues, but also as part of automated education and training, such as e-learning. But there remains an inherent tension between ML performance (for example, predictive accuracy) as well as ideas of fairness and explainability. Often the highest-performing methods (for example, DL) are the least explainable, and the most explainable (for example, decision trees) are the least accurate and do not take into account the needs of the user.

Respective of formalisms and computational methods, an important criteria driving development is to ask "do these explanations make something clear?" DL systems are opaque and do not fully handle desired aspects of explanation to make them humanly comprehensible, which is the ability, in this case of ML algorithm to represent what is learned in a human understandable fashion. As noted, technical views may provide an answer to "how" the explanation was arrived at in steps and which rules or features were involved, but not the justifying and clarifying "why" of a satisfactory explanation. If, for example, a tree or hierarchical structure is involved in an explanation process we might get more of a "why" understanding with the possibility of drilling down and browsing a decision tree, having a focal point of attention on critical information or having the option of displaying a graphic representation that is human understandable. An example would be if a vehicle controller AI system for driving, based on visual sensing of objects could provide commonsense explanations (Persaud *et al*, 2017). Using internal commands a system may describe itself spatially as "moving forward", while a human description is the more functional and just one of "driving down the street." For explaining a lane change the system says, "because there are no other cars in my lane" while the human explanation is informative in another way "because the lane is clear." These are similar but "clear" is a more comprehensive idea of a situation which might include construction, tree litter *etc.* (Tandon *et al.* 2018). A comprehendible explanation includes coherent pieces of information, more or less directly interpretable in natural language, and might relate quantitative ("no cars" and qualitative concepts ("near my lane") in an integrated fashion.

It is important to note that under the influence of modern ML and Deep Learning (DL) models both CSKR and smart system explanations have recently been developing alongside these efforts and provide mutual support by co-developing deep explanations. These amount to modified or hybrid DL techniques that learn more explainable and CSKR features or representations or that feed into explanation generation facilities.

An area where we might see this developed is in the ability of DL-based applications to describe images (Geman, *et al.* 2015). This might be considered as one element of a visual Turing Test-like application and involve question- answering based on real-world images, such as detecting and localizing instances of objects and relationships among the objects in a scene. Some commonsense-making involved localizing questions[x] posed might include the following:

- What is Person1 carrying?
- What is Person2 placing on the table?
- Is Person3 standing behind a desk?
- Is Vehicle 1 moving?

This spate of recent work, reflecting the ability of ML systems to learn and answer questions about visual information and even text, has led to more distinctions being made about CSKR in support of robustness of the many ML applications which are increasingly thought of as mature enough to use for some ordinary tasks. Visual recognition is one of these, and supporting research approaches generate image captions to train a second deep network that can in turn generate explanations without explicitly identifying the original network's semantic features. This work continues but Shah *et al* (2019) suggests that some current ML applications are not robust as simple alternative NL syntactic formulations that lead to different answers. For example, if a system is asked "What is in the basket" and "What is contained in the basket" (or "what can be seen inside the basket") we get very different answers. Humans understand these as having similar commonsense meanings, but ML systems may have learned something different. And we may not know what they have learned and thus any direct explanation may be unsatisfactory for a user.

An obvious problem is that DL using a combination of efficient learning algorithms working over huge parametric space by themselves, are complex black-boxes in nature (Castelvecchi, 2016). For example, in a large knowledge graph measurements like nearest neighbors cannot be decomposed and/or the number of variables is so high that the user has to rely on mathematical and statistical tools to analyze the model. So while these approaches allow powerful predictions, their raw outputs cannot easily be directly explained and post hoc efforts are sometimes used. Consider the capability people have to distinguish the visual modality expressing a simply observed property like color or what seems like some simple relation like part. These afford common-sense and practical implications like "shiny things imply smoothness and so less friction". Distinctions like "smoothness" can play a role in transfer of training to new areas. Research now reliably shows the value of transfer training/learning such as with NELL. Transfer is enabled by pre-training a neural network model working on a known task, say image recognition using stored images from a general source like ImageNet. The resulting trained neural network (with an implied "model") is aimed for use with new, but related and purpose-specific models. What makes transfer difficult is finding training data that can provide a base to transfer for multiple types of scenarios and new situations of interest. There remain problems of representativeness and the selection of the typical to some generalizations such "shiny surfaces are typically hard, but some are not". There is also the problem of perspective. Imagine that we have in an image, the moon in the sky and a squirrel under a tree. They may seem the same size, but we know from common experience that they are at different distances and thus only appear to have a similar size. This is not something learned by a regular NN application, but it would be good to acquire this type of CSKR to allow this understanding.

Summing up Findings, Directions and Future Work

It seems clear that both CSKR and explanation remain important topics as part of AI research and its surging branch of ML. Further they can be mutually supportive, although explanation may be the more active area of diverse work just now. A guiding idea is that a truly explainable model should not have such knowledge gaps that users are left to generate different interpretations depending on their background knowledge. Having a suitable store of CSK can help an intelligent system produce explanations including natural language forms combining CSKR and human-understandable features (Bennetot, 2019).

For future direction five areas are noted:

1. Challenges in developing an adequate base of CSK
2. Situational and contextual understanding
3. Deep learning and dynamic situations
4. Interactions with humans
5. The need for a common, enhanced ontology engineering practice.

Adequate Knowledge: Providing a suitable base of CSK remains a broad, deep, and some say a largely unbounded problem. It seems generally true that one master ontology will not suffice for either specific domains or CSK and that a range of ontologies will be needed for an adequate CSK. Single ontologies are not likely to be suitable as work expands and more contexts are encountered. This will require multiple ontologies and/or a range of MTs as in Cyc. Big Knowledge, with its heterogeneity and depth complexity, may be as much of a problem as Big Data especially if we are leveraging heterogeneous, noisy, and conflicting data to create CSKR and explanations (Pauleen and Wang 2017). Various approaches do exist for different forms of CSKR, but the integration of these as well as ontologies with different content is still challenging. Linked data have a simplified view of KBs as a set of linked sentence-like assertions. However, integration of these requires some degree of background knowledge to understand the underlying assertions expressed in natural language labels. It is hard to imagine that major integration challenges from various forms with varying degrees of formality can be avoided. The ontology experience is that as a model of the real world we need to select some part of reality based on interest and conceptualization of that interest. Differences of selection and interpretation are impossible to avoid and it can be expected that different external factors will generate different contexts for any intelligent agent doing the selections and interpretation needed as part of a domain explanation.

The work such as Yi and Michael Gunginer, 2018 (Gunninger, 2018) suggests some coordinated set of ontologies that might be needed to support something as reasonable and focused as a Physical Embodied Turing Test. These include several aspects of intelligence, such as perception, reasoning, and action. Grunninger's suite (Gruninger, 2019), called PRAxIS (Perception, Reasoning, and Action across Intelligent Systems) with the following components:

- Solid Physical Objects (SoPhOs)
- Occupy (Location - Occupation is a relation between a physical body and a spatial region)
- Process Specification Language (PSL)
- Processes for Solid Physical Objects (ProSPerO)
- Ontologies for Video (OVid)
- Foundational Ontologies for Units of measure (FOUnt)

It is worth mentioning that ontologies like SoPhOs might emulate the intuitive physics of child cognition for objects while an "Occupy" concept provides notions of location and place used for spatial navigation. While this remains an early effort it does illustrate some of the diverse types of CSKR that need to be formalized. Yet there exists a range of strategies that could be employed to make progress on both the CSKR challenge and in its use to enhance explanations. In the sub-sections below, some of the remaining explanation and CSKR issues are further illustrated arising from some old problems that may affect more on the relatively newer challenges raised by ML and DL approaches.

Situational and contextual understanding: More complex tasks will involve greater situational understanding[xi]. These include situations where important things are unseen, but implied in a picture as part of the larger or assumed context such as exist in environmental or ecological settings with many dependencies. An example offered by Niket Tandon (Tandon *et al.*, 2018) involves the implication of a directional arrow in a diagram of food web which intentionally communicates "consumes" to a human (a frog consumes bugs). The problem for modern learning oriented systems is that they are unlikely to have arrows used visually this way enough to generalize to a "consumes" meaning. To a human this is background knowledge.

Alas, it remains a hard problem to engineer all such knowledge or acquire it in an automated fashion. Indeed, since their inception, both explanatory systems and commonsense R & D have proven to involve implied, hard problems addressed by natural biological evolutions over a long period of time: such as the ideas of effective communication, consensus reality, background knowledge, notions of causality, and rationales. These allow the handling of things like focus and scale that is a known problem in visual identification. In a lake scene with a duck a ML vision system may see water features like dark spots as objects. In this case there seems a need for a model of the situation and for what is the focus of attention – a duck object. Some use of commonsense as part of model-based explanations might help during model debugging and decision making to correct apparently unreasonable predictions.

Such problems seem simple only because these are ubiquitous in everyday thinking, speaking, and perceiving as part of ordinary human interaction with the world. And this knowledge and reasoning seems easily captured because it is commonly available to the overwhelming majority of people, and manifest in human children's behavior by the age of three or five.

Deep Learning and Dynamic Situations Generally, the current state of the art for ML suggests that deep learning can provide some explanations of what they identify in simple visual datasets such as Visual Query Answering (VQA) and CLEVR. They can answer questions like "What is the man riding on?" in response to an image such as the one in Figure 3.

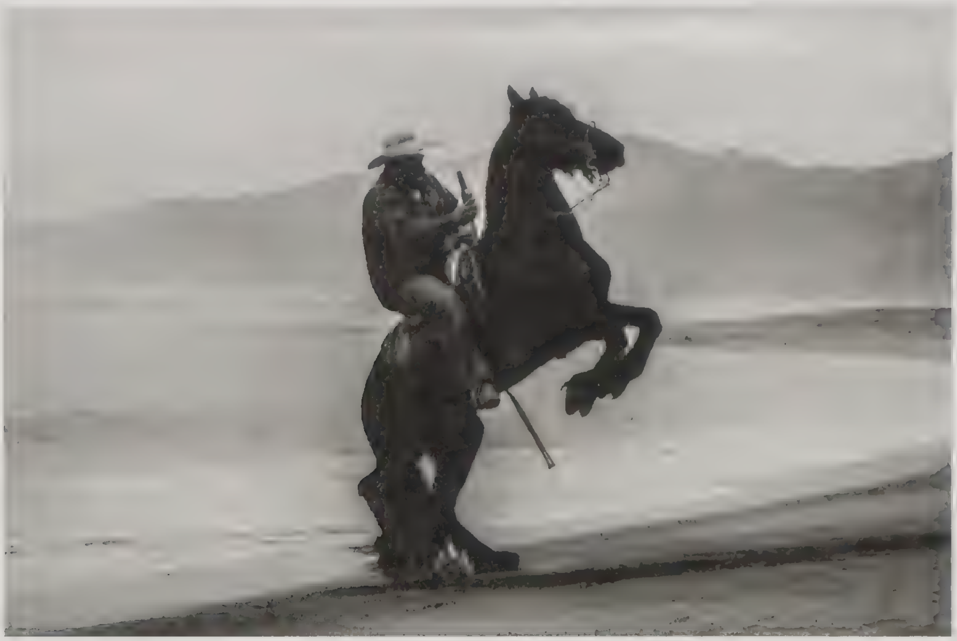


Figure 3: Example of image for ML processing

Whereas, commonsense knowledge is more important when the visual compositions are more dynamic and involve multiple objects and agents typical of say a cattle roundup. For dynamic and other situations further advanced intelligent system evolution needs to consider other features that may be supportive. This is true even in leaning-oriented systems like NELL which extract information from sentences. Because of things like contextual relations there remain many problems with unsophisticated textual understanding. Examples are the implications and scope of negations and what is entailed.[xii]

Beyond negations there may be many situations one needs to understand – for example, "what exactly is happening in this ecological view?" This is challenging because a naive, start from scratch computational system, has to track everything involved in a situation or event. This may involve a long series of events with many objects and agents as in an ecological example or a food chain. Previously discussed situational complexity is also evident in visualizing a routine procedural play in basketball even as simple as a completed or missed dunk (Mishra *et al.*, 2019). Images of a dunk attempt can be described by three NL sentences: "He charges forward. And made a great leap. He made a basket." These sentences may be understood in terms of some underlying state-action-changes with a sequence of actions such as running and jumping, but there are also implied states as follows:

- The ball is in his hands. (not actually said, but seen and important for the play)
- The player is in the air. (implied by the leap)
- The ball is in the hoop. (technically how a basket is made)

We can represent the location of things in the three sentences above like this:

- Location (ball) = player's hand
- Location (player) = air
- Location (ball) = hoop (after Tanden, 2019)

These all fit into a coherent action with the context of a basketball script that we know, and thus humans can focus on the fact that the location of the ball at the end of the jump is a key result. CSKR about bodily capabilities apply here (Can I reach that hoop by jumping?) On other hand, as shown by Tandon *et al.* (2018), it is expensive to develop a large enough training set for such CSKR of activities, and the resulting state-action-change models have so many possible inferred candidate structures (*e.g.*, is the ball still in his hand? Maybe it was destroyed) so that common events can evoke an NP-complete problem. Without sufficient data (remember it is costly to construct), the model can produce what one would consider to be absurd, unrealistic choices based on commonsense experience such as the player being in the hoop.

A solution is to have a commonsense aware system that constrains the search for plausible event sequences. This is possible with the design and application of a handful of universally applicable rules. And some constraining ruling can be derived from existing ontologies. For example these constraints seem reasonable based on commonsense:

1. An entity must exist before it can be moved or destroyed. (certainly not likely in basketball)
2. An entity cannot be created if it already exists.
3. A tennis player is located at a tennis court.

In the work discussed by Niket Tanden (2018) these constraints were directly derivable from the Suggested Upper Merged Ontology (SUMO) rules such as: MakingFn, DestructionFn, MotionFn. This provides preliminary evidence that ontologies, even early ones such as SUMO, could be good guides for producing a handful of generic hard constraints in new domains.

One might ask, "How much help do these constraints provide?" The answer is that CSK-based search improves precision by nearly 30% over State-Of-The-Art DL efforts which include Recurrent Entity Networks (EntNet) , Query Reduction Networks (QRN) , and ProGlobal (Tanden *et al*, 2018).

Humans in the Loop While we do seem close to AI systems that will do common tasks such as driving a car or give advice on common tasks like eating it remains a challenge that such everyday tasks exhibit robust CSK and reasonable reasoning in order to be trusted. Monitoring the reasonableness and safety of automated actions, like driving in dynamic or even novel situations, illustrate a rapidly approaching but still challenging commonsense service capability. As intelligent agents become more autonomous, sophisticated, and prevalent, it becomes increasingly important that their knowledge become more complete and that humans be able to interact with them effectively to answer such questions as "why did you (my self-driving vehicle) take an unfamiliar turn?"

We need humans in the loop and allow dynamic interactions with intelligent agents. It is widely agreed that we need to enable humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners (Arrieta, 2020). Defining a successful application and its explanations remains relative to its audiences and their understanding. This is a bit of a psychological task so we can't expect system designers and engineers to solve this without help (Mueller *et al*, 2019). But engineers can understand that human interactions and reactions to poor explanations can help to detect, and thus, correct things like bias in the training dataset or in system reasoning.

Current AI systems are good at recognizing objects, but can't explain what they see in ways understandable to and somewhat explainable by laymen. Nor can systems read a textbook and understand the questions in the back of the book which leads researchers to conclude they are devoid of common sense. We agree, as DARPA's Machine Common Sense (MCS) proposal put it that the lack of a common sense is "perhaps the most significant barrier" between the focus of AI applications today (such as previously discussed), and the human-like systems we dream of. And at least one of the areas that such an ability would play is with useful explanations. It may also be true, as NELL researchers argue, that we will never produce true NL understanding systems, until we have systems that react to arbitrary sentences with "I knew that, or didn't know and accept or disagree because X".

Better Methods for Engineering CSKR and Explanation: It is also worth noting that as explanation and CSKR research converge there is a need to develop a common, enhanced ontology engineering practice. As we arrive at a more focused understanding of CSKR there will be a need for this convergence to be incorporated into common ontological engineering practices. For efforts like CSK base building this should include guidance and best practices for the extraction of rules from extant, quality ontologies. A particular task is evaluating the quality of knowledge, both CSK and domain knowledge extracted from text. If knowledge is extracted from text and online information building of CSK will require methods to clean, refine and organize them. It is not as simple as saying that a system provides an exact match of words to what a human might produce given the many ways that meaning may be expressed. And it is costly to test system generated explanations or even captions against human ones due to the human cost.

One interesting research approach is to train a system to distinguish human and ML/DL system generated captions (for images *etc.*). After training one can use the resulting learned distinguished systems to critique the quality of the ML/DL generated labels.

In some cases, and increasingly so, a variety of CSK/information extracted is aligned (*e.g.* some information converges from different sources) by means of an extant (hopefully of high quality) ontology and perhaps several. This means that some aspect of the knowledge in the ontologies provides an interpretive or validating activity for the structuring involved in building artifacts like KGs. Knowledge graph gaps can also be filled in by internal processes looking for such things as consistency with common ideas as well as from external processes which adds information from human and/or automated sources. KG building efforts, which started employing sources like Freebase's data as a "gold standard" to evaluate data in DBpedia which in turn is used to populate a KG, are moving on to augmentation from text sources. In this light we can again note that a key requirement for validated table quality of knowledge involves the ability to trace back from a KB to the original source documents (such as LinkedData) and if filled in, from other sources such as humans to make it understandable or trustworthy. It is useful to note that this process of building such popular artifacts as KGs clearly shows that they are not equivalent in quality to supporting ontologies. In general there is some confusion in equating the quality of extracted information from text, KGs, KBs, the inherent knowledge in DL systems and ontologies.

But all such efforts are very probably going to rely on the assistance of new as yet undeveloped tools. In light of this future work we will need to refine a suite of tools and technologies to make the lifecycle of commonsense KBes easier and faster to build.

A successfully engineered intelligent system would be more of an "Associate Systems" with which users dialog with and over time get satisfactory answers because they include a capability to adaptively learn user knowledge and goals and are accountable for doing so over time. This is, of course, commonly true for human associates. The idea here is to mirror the user's mental model including some idea of commonsense, which becomes one of the main building block of intelligible human–machine interactions. Such focused, good, fair explanations may use natural language understanding to be part of a conversational dialogue human-computer interaction (HCI) in which the system uses previous knowledge of user (audience) knowledge and goals to discuss output explanations.

In such associate systems an issue will be the focus of attention. As part of common experience focus is an important element of explanations and commonsense assumptions and presumptions in a knowledge store play an important role in focus point. Indeed the ability to focus on relevant points may be part of the way a system competence is judged. But good focus has many potential dimensions and can involve judging and evaluating technical factors such as ethicality, fairness, and, where relevant, legality along with various roles such as relational, processual role, and social roles. These will all be important aspects of advanced AI applications. An example of this is that the role of legal advice is different in the context of a banking activity as opposed to lying under oath.

Acknowledgements

I thank Todd Schneider for a friendly review and suggestions.

References

1. Arrieta, Alejandro Barredo, *et al*. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* **58** (2020): 82-115.
2. Baclawski, K., Bennett, M., Berg-Cross, G., Casanave, C., Fritzsche, D., Luciano, J. & Sriram, R. D. (2018). Ontology Summit 2018 Communiqué: Contexts in context. *Applied Ontology*, **13**(3), 181-200.
3. A. Bennetot, J.-L. Laurent, R. Chatila, N. D´ıaz-Rodr´ıguez, Towards explainable neural-symbolic visual reasoning, in: NeSy Workshop IJCAI 2019, Macau, China, 2019.
4. Bundy, Alan. "A critique of proof planning." *Computational Logic: Logic Programming and Beyond*. Springer, Berlin, Heidelberg, 2002. 160-177.
5. Castelvecchi, D. Can we open the black box of AI? *Nature News* 538 (7623) (2016) 20.
6. Chakraborty, Supriyo, et al. "Interpretability of deep learning models: a survey of results." *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 2017.
7. Cohen, Robin, et al. "Trusted AI and the contribution of trust modeling in multiagent systems." *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2019.
8. Clancey, William J. "The epistemology of a rule-based expert system—a framework for explanation." *Artificial intelligence* 20.3 (1983): 215-251.
9. Darlington, Keith. "Aspects of intelligent systems explanation." *Universal Journal of Control and Automation* 1.2 (2013): 40-51.

10. Davis, Ernest, and Gary Marcus. "Commonsense reasoning and commonsense knowledge in artificial intelligence." *Communications of the ACM* 58.9 (2015): 92-103.

11. Fox, Maria, Derek Long, and Daniele Magazzeni. "Explainable planning." arXiv preprint arXiv:1709.10256 (2017).

12. Geman, Donald, *et al.* "Visual turing test for computer vision systems." *Proceedings of the National Academy of Sciences* 112.12 (2015): 3618-3623.

13. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016, http://www.deeplearningbook.org

14. Grosof, Benjamin, *et al.* "Automated decision support for financial regulatory/policy compliance, using textual rulelog." *Financial Times* (2014).

15. Gunning, David. "Machine common sense concept paper." *arXiv preprint arXiv:1810.07528* (2018).

16. Gruninger, Michael, Ontologies for the Physical Turing, Ontology Summit 2019 January, 2019 https://s3.amazonaws.com/ontologforum/OntologySummit2019_Commonsense_summit-physical-turing.pdf

17. Havasi, Catherine. "Reflections on Structured Common Sense in an Era of Machine Learning." *Proceedings of the 10th International Conference on Knowledge Capture*. 2019.

18. Holland, John H. "Escaping brittleness." *Proceedings Second International Workshop on Machine Learning*. 1983.

19. Kang, Dongyeop, *et al.* "Bridging Knowledge Gaps in Neural Entailment via Symbolic Models." *arXiv preprint arXiv:1808.09333* (2018).

20. Langley, Pat, *et al.* "Explainable Agency for Intelligent Autonomous Systems." AAAI. 2017.

21. Langlotz, C., and Shortliffe, E. Adapting a Consultation System to Critique User Plans. In Coombs, M. (Editor). Developments in Expert Systems. Academic Press, London. 1984.

22. Letham, B.; Rudin, C.; McCormick, T. H.; and Madigan, D. 2015. Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. Annals of Applied Statistics

23. Lenat, Douglas B., Mayank Prakash, and Mary Shepherd. "CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks." *AI magazine* 6.4 (1985): 65-65.

24. Lenat, Douglas B., and Ramanathan V. Guha. *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc., 1989.

25. Lifschitz, Vladimir. "Formalizing Common Sense: Papers by John McCarthy." Ablex, Norwood, NJ (1990).

26. Lipton, Z.C.: The mythos of model interpretability. Workshop on Human Interpretability in Machine Learning (2016)

27. McCarthy, John. *Programs with common sense*. RLE and MIT computation center, 1960

28. McGuinness, Deborah L., and Paulo Pinheiro Da Silva. "Infrastructure for web explanations." *International Semantic Web Conference*. Springer, Berlin, Heidelberg, 2003.

29. Melis, Erica. "AI-techniques in proof planning." *The planner* 1.2 (1998): 3.

30. Miller, Tim. "Explanation in artificial intelligence: insights from the social sciences." arXiv preprint arXiv:1706.07269 (2017).

31. Minsky, Marvin. "Commonsense-based interfaces." *Communications of the ACM* 43.8 (2000): 66-73.

32. Minsky, Marvin L., Push Singh, and Aaron Sloman. "The St. Thomas common sense symposium: designing architectures for human-level intelligence." *Ai Magazine* 25.2 (2004): 113-113.

33. Mishra, Bhavana Dalvi, *et al*. "Everything Happens for a Reason: Discovering the Purpose of Actions in Procedural Text." *arXiv preprint arXiv:1909.04745* (2019).

34. Mitchell, Tom, *et al*. "Never-ending learning." *Communications of the ACM* 61.5 (2018): 103-115.

35. Moreau, Luc. "The foundations for provenance on the web." *Foundations and Trends® in Web Science* 2.2–3 (2010): 99-241.

36. Mueller, Shane T., *et al*. "Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI." *arXiv preprint arXiv:1902.01876* (2019).

37. Ontology Summit, https://ontologforum.org/index.php/OntologySummit2019 2019.

38. Ortiz Jr, Charles L. "Why we need a physically embodied Turing test and what it might look like." *AI magazine* 37.1 (2016): 55-62.

39. Pauleen, David J., and William YC Wang. "Does big data mean big knowledge? KM perspectives on big data and analytics." *Journal of Knowledge Management* (2017).

40. Panton, Kathy, *et al*. "Common sense reasoning from Cyc to intelligent assistant." *Ambient Intelligence in Everyday Life*. Springer, Berlin, Heidelberg, 2006. 1-31.

41. Persaud, Priya, Aparna S. Varde, and Stefan Robila. "Enhancing autonomous vehicles with commonsense: Smart mobility in smart cities." *2017 IEEE 29th*

International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2017.

42. Singh, Push. "The open mind common sense project." *KurzweilAI. net* (2002a).
43. Singh, Push. "The public acquisition of commonsense knowledge." *Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*. 2002b.
44. Shah, Meet, *et al*. "Cycle-consistency for robust visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
45. Suchman, L. A. (1987). Plans and situated actions: The problem of human-machine communication. Cambridge university press.
46. Swartout, William R., and Stephen W. Smoliar. "Explanation: A source of guidance for knowledge representation." *Knowledge Representation and Organization in Machine Learning*. Springer, Berlin, Heidelberg, 1989. 1-16.
47. Tandon, Niket, Aparna S. Varde, and Gerard de Melo. "Commonsense knowledge in machine intelligence." *ACM SIGMOD Record* 46.4 (2018): 49-52.
48. Tandon, Niket, *et al*. "Reasoning about actions and state changes by injecting commonsense knowledge." *arXiv preprint arXiv:1808.10012* (2018).
49. Tandom, Niket   Commonsense for Deep Learning, presented at Ontology Summit, 2019 http://people.mpi-inf.mpg.de/~ntandon/presentations/ontology-summit-2019/ontology-summit2019-niket-tandon.pdf
50. Von Rueden, Laura, *et al*. "Informed machine learning–towards a taxonomy of explicit integration of knowledge into machine learning." *Learning* 18 (2019): 19-20.
51. Walton, Douglas. *Abductive reasoning*. University of Alabama Press, 2014.
52. Yi, R. U., and Michael GR UNINGER. "What's the Damage? Abnormality in Solid Physical Objects." (2018).
53. Yuan, Changhe, Heejin Lim, and Tsai-Ching Lu. "Most Relevant Explanation in Bayesian Networks." *Journal Of Artificial Intelligence Research* (2011).

[i] The idea of intelligent systems covers a broad range of software technologies from simple heuristic and rule-based systems emulating human expertise with symbolic processing, to more recent neural network and machine learning technologies..

[ii] In "Programs with Common Sense" McCarthy (1960) described 3 tactical ways for early AI to proceed which includes common sense understanding and imitating the human central nervous system, which to a degree NN

systems do study human cognition or "understand the common sense world in which people achieve their goals."

iii In a narrow, logical and technical sense the "Gold Standard" concept of explanation is such a faithful, deductive proof done using a formal knowledge representation (KR).

iv These descended from J. McCarthy's tradition of treating contexts as formal objects over which one can quantify and express first-order properties.

v For example, "fact sentence F41 was drawn from document D73, at URL U84, in section 22, line 18." That kind of explanation is valuable and allows follow up.

vi Obviously a machine capability for a basic level of human-like commonsense would enable more effective communication and collaborate with their human partners.

vii As Andrej Karpathy put it, "I don't have to actually experience crashing my car into a wall a few hundred times before I slowly start avoiding to do so."

viii Reasoning is also applied for consistency checking and removing inconsistent axioms as in other knowledge graph (KG) generation efforts.

ix Knowledge reuse and transfer is an important issue in making such systems scalable.

x Broadly we might conceptualize this as a type of sensemaking in which an intelligent system that needs to analyze and interpret sensor or data input benefits from a CSKR service providing help it interpret and understand real world situations.

xi Some of these still unsolved contextual issues were discussed as part of the Ontology Summit 2018 on contexts (Baclawski *et al.*, 2018).

xii Kang *et al* (2018) showed the problems of what is concluded based on textual entailment with sentences from the Stanford Knowledge Language Inference set with sentences like "The dog did not eat all of the chickens."

BIO

Gary Berg-Cross is a cognitive psychologist (PhD, SUNY–Stony Brook) whose professional life included teaching and R&D in applied data & knowledge engineering, collaboration, and AI research. A board member of the Ontolog Forum he co-chaired the Research Data Alliance work-group on Data Foundations and Terminology. Major thrusts of his work include reusable knowledge, vocabularies, and semantic interoperability achieved through semantic analysis, formalization, capture in knowledge tools, and access through repositories.

Applied Ontologies for Global Health Surveillance and Pandemic Intelligence

Christopher J. O. Baker[1, 6], Mohammad Sadnan Al Manir[5], Jon Hael Brenas[3], Kate Zinszer[4], Arash Shaban-Nejad[2,*]

[1]Department of Computer Science, University of New Brunswick, Saint John, NB, Canada
[2] The University of Tennessee Health Science Center - Oak Ridge National Laboratory (UTHSC-ORNL) Center for Biomedical Informatics, Department of Pediatrics, College of Medicine, Memphis, TN, United States
[3] Big Data Institute - Nuffield Department of Medicine, University of Oxford, Oxford OX3 7LF, U.K.
[4] School of Public Health, University of Montreal, Montréal, Québec, Canada.
[5] Public Health Sciences, University of Virginia, Charlottesville, VA, USA.
[6] IPSNP Computing Inc, Saint John, NB, Canada

Abstract

Global health surveillance and pandemic intelligence rely on the systematic collection and integration of data from diverse distributed and heterogeneous sources at various levels of granularity. These sources include data from multiple disciplines represented in different formats, languages, and structures posing significant integration challenges. This article provides an overview of challenges in data driven surveillance. Using Malaria surveillance as a use case we highlight the contribution made by emerging semantic data federation technologies that offer enhanced interoperability, interpretability, and explainability through the adoption of ontologies. The paper concludes with a focus on the relevance of these technologies for ongoing pandemic preparedness initiatives.

Introduction

WHEREAS HEALTH SURVEILLANCE has always been an essential activity; the recent global health crisis caused by COVID-19 has highlighted our dependence on bespoke surveillance infrastructures that provide support for decisions of considerable gravity. Core to the success of these endeavors is the coordination of multiple data sources; however, many challenges related to data management remain unresolved and limit the insight we can gain from data-driven surveillance. These challenges stem from incumbent infrastructures where collected data are stored in distributed heterogeneous siloed information systems without enabling metadata or standardization, posing interoperability and data integration challenges [1]. The absence of interoperability results in poor coordination between surveillance systems which are known to be rigid [2], leading stakeholders to have limited confidence in any insights derived from the aggregated datasets [3]. These barriers delay the generation of key insights that are needed to support decision making and to plan appropriate and timely interventions. Specifically for decision-makers, there is also an intelligibility problem, which requires a transparent understanding of sources of data and data processing activities to generate trustable insights. Here, surveillance systems need to enable explanations based on provenance annotations. The requirement for explainability becomes increasingly important given the diversity of data sources being harvested in disease surveillance. Indeed, an increasingly broad array of stakeholders now contributes vast sensor data from new devices and text from social media platforms, augmenting the complexity of sources and data structures.

In the case of malaria it has been reported that few information systems can comprehensively collect, store, analyze data, and provide feedback based on real-time information [4]. Additionally, concurrency control issues can emerge, where data entered from field stations are recorded centrally but are not immediately reflected at the field level [5-7]. Developers and users trying to gain remote access to data with web services have also identified that they are inflexible for reuse and there is a lack of standardization. In fully deployed systems the resolution of spatiotemporal data [8] is often limited in scope, *e.g.* not to the level of individual households, nor is it possible to extrapolate trends across geographic borders. Consequently, the types of surveillance queries that can be run to derive actionable knowledge are relatively rigid and reporting tools cannot support ad hoc queries without significant redevelopment. A recent WHO technical strategy document [9] stated that malaria surveillance mechanisms designed to facilitate interoperable data integration from distributed data silos are lacking.

These types of challenges, interoperability, interpretability, and explainability, have motivated computer scientists to consider how the provisional data for a wide range of stakeholders lead to the introduction of a set of guidelines seeking to ensure that published data are findable, accessible, interoperable, and reusable [105], albeit no specific technical solutions are proposed or mandated.

Broadly speaking, interoperability can be understood as the provision of common interfaces between divergent computer systems to ensure seamless access to data and services. It can include enriching data or services with context, unambiguous meaning, and provenance in a standardized syntax or format in support of data exchange and reuse. Practically, it means multiple users can mutually access and reuse data without having to store it locally or reformat the data on import, knowing that the integrity and meaning (semantics) of the data have been maintained since its creation to its application and reuse. Provisioning interoperability is addressed by the mapping of data to standardized vocabularies or terms in ontologies [11]. Ontologies are representations of domain knowledge using concepts, relations, and complex logical rules or axioms. Interoperability can be achieved by mapping data sets to the same ontology terms or vocabularies to ensure they can be regarded as having the same meanings.

Such representations of domain knowledge and metadata also underpin the explainability of decision support in a surveillance platform where the reasoning behind the decision can be made transparent to the end-users. In order to achieve this transparency, the data resources supporting a decision need to be verified, preferably in real-time. One method to verify the resources used to achieve the decision is by resolving all their locations from their URIs. Once resolved, they lead to the locations of the actual content or the associated metadata based on the access policy. In a service-based surveillance system [12], resources typically include the input data, and services, workflows, software, and scripts which produce the output data. Various approaches to verification also referred to as *provenance*, have been investigated and tried in the context of data [13], software [14], and workflow [15, 16]. A standard for vocabularies to represent provenance information in a formalized way is the W3C PROV Ontology (PROV-O) [17] which can be used as a vehicle to formalize explainability.

Disease Surveillance Ontologies

To be effective disease surveillance ontologies [18, 19, 20, 21] must cover a range of perspectives including vector biology, etiology, transmission, pathogenesis, diagnosis, prevention, and treatment. Depending on the intended use of the ontology, these perspectives have been represented in different ways leading to ontologies of different maturity, expressiveness, and fit-for-purpose. Whereas a detailed review of these would be beyond the scope of this article, here we briefly review a small sample of ontologies primarily to illustrate the breadth and scope of the domain knowledge that needs to be represented to support surveillance interoperable surveillance infrastructures.

For vector-borne diseases, the Vector Surveillance and Management Ontology (VSMO) [18] focuses on arthropod vectors and vector-borne pathogens specific to domestic animals and humans as well as the corresponding surveillance data management systems, or decision support systems. A core feature of the VSMO is the vector relation, linking arthropod species (vectors) to pathogenic microorganisms.

The Infectious Disease Ontology-Malaria (IDOMAL) [19] is an extension of the infectious disease ontology (IDO) [22] that includes broad coverage of malaria including clinical manifestations, therapeutic approaches, epidemiology, vector biology, and insecticide resistance (IR). Vector physiology is modeled from the perspective of processes related to transmission, particularly interactions between the vector and the vertebrate host of Plasmodium, as well as the vector and Plasmodium itself. This extends to behavioral parameters such as host-seeking or blood meal-related processes.

Mosquito Insecticide Resistance Ontology is an application ontology for entities related to insecticide resistance in mosquitoes. MIRO [20] was designed to support IRbase, a dedicated resource for storing data on insecticide resistance in mosquito populations. It focuses on archiving information on geolocations, mosquito populations as the main vectors of diseases (dengue, filariasis, malaria, yellow fever), types of assays performed to assess types, and levels of insecticide resistance to support the design of interventions.

Animal Health Surveillance Ontology (AHSO) [21] is an excellent example of this, designed to support decision making based on data that were collected for alternative purposes, including clinical records, laboratory findings, or slaughter inspection data. In this case herd information is collected along the entire cycle of animal production, even in the absence of disease events. The targeted surveillance analytics makes use of all recorded observations such as a disease occurrence, births, or product yield (dairy or livestock). Ontologies that support surveillance analyses and automation of tasks translating data into actionable information must accommodate the production system, the nature of observation, and the context in which the data was recorded. AHSO represents definitions of syndromes and models observations that relate to health events at specific moments in time but not the actual health events. The ontology has three main levels: sample, observations, and observational context, for instance, a clinical observation, or surveillance sampling activity. A health event is modeled as an abstract concept with undefined boundaries in time, space, or population units. It is assumed that several observations are derived from a health event and recorded in one or more databases [21].

In general the adoption of ontology models by users other than the ontologists that built them can depend on many factors related to target goals and purpose. The design of an ontology, so that it is fit for purpose, can vary greatly, and ontologies designed for different purposes by different communities generally result in different ontologies, both in terms of scope and structure, which can occur for even the same subject matter.

Deciding whether a domain ontology built for any of these purposes and auxiliary activities can be reused is daunting and requires considerable expertise and time-investment. For these reasons the reuse and adoption of any given ontology is generally a slow process requiring a full evaluation of what aspects of an ontology can be useful in a new context. Sometimes it is easier to start a new ontology and then subsequently do a mapping to any related ontologies that may exist. Where ontologies or parts of ontologies have been imported to new ontologies this can be identified by reviewing ontology files for imports with different URLs to reveal which parts of an ontology are from other conceptualizations. Some studies have sought to elucidate such artifacts [39] with varying degrees of success.

To further highlight the challenges of reuse we list here some common motivations for ontology development: (i) to share a common understanding of information; (ii) to enable reuse of knowledge through explicit representation of knowledge and formal reasoning; (iii) the derivation of further insights in a domain aka knowledge discovery; (iv) to enable reasoning and quality control (*i.e.* revealing inconsistencies and insatisfiabilities); and (v) to improve reusability, maintenance, versioning, and change management. The precise manifestation of these goals can be quite technical and here we point primarily to the classification of ontologies to identify inconsistencies in a knowledge representation about a domain

[23], classification of instance data based on formal axioms or rules in a domain ontology [24, 25], authoring of knowledge graphs using ontologies as a reference model [26], ontology-based data access [27], and the use of ontology terms for authoring of semantic web services [28]. All of these are specific cases that leverage more than the primary formal conceptualization of a domain built for the sake of knowledge sharing. Overall the maturity of the model and its design and purpose are limiting factors for reuse. One good example of an ontology that has been well cited and adopted is the Semantic science Integrated Ontology (SIO) [29], which provides a simple, integrated upper-level ontology (types, relations) for consistent knowledge representation across physical, processual, and informational entities. It is broadly adopted in the life sciences because of its design and relevance to many use cases.

Disease Surveillance tasks

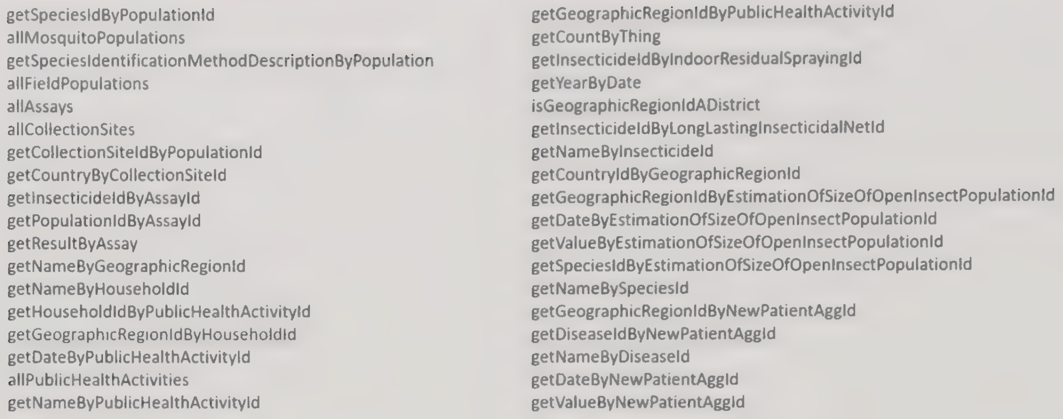
Surveillance is an activity that involves a series of tasks; monitoring and harvesting of data, analysis of data to review the disease trends followed by the design of targeted interventions, their implementation, and evaluation. To be effective surveillance is an iterative lifecycle of tasks and activities with the goal of harvesting actionable data. What makes it challenging is that surveillance practitioners need to obtain custom views of target data and decision parameters in a timely and non-arbitrary manner. Often there is a lack of understanding of such parameters, such as a reproduction number representing a disease's ability to spread [30], which can lead to ill-informed decisions and public health interventions by different stakeholders. Determining the effectiveness of interventions, by combining reporting and cross-checking with multiple indicators and data sources, is essential. In particular there is a need to support surveillance practitioners with *ad hoc* querying over integrated data. Existing infrastructures are limited to delivering information on a fixed set of defined parameters. Agility is essential, and surveillance systems relying on slow to deploy information gathering pipelines lacking interoperability are insufficient, particularly when requirements shift *e.g.* to understanding demographics of infections, in addition to overall infection rates.

In light of these new requirements for surveillance systems, a new generation of surveillance platforms is emerging that can address the provision of agility and ad hoc querying for non-technical users.

Surveillance systems need to be able to discover and select data sets that contribute to a given line of inquiry from a registry of available data services and data transformation services. The essential tasks that need to be supported are: (i) the use of precise and formal semantics to describe input and output of data services to ensure they are rapidly discoverable at the time of the query; (ii) the provision of search engines that can understand such semantic descriptions; (iii) the provision of interoperability between services to ensure complex workflows needed for retrieving and transforming data are uninterrupted; and (iv) the provision of intelligible query composition tools that are readily explainable to novice users.

For over a decade these design requirements have been core to semantic web services frameworks which have been recently deployed in surveillance use cases. In [28, 33, 34] the Semantics, Interoperability, and Evolution for Malaria Analytics (SIEMA) platform was deployed for use in malaria surveillance based on semantic data federation. SIEMA's objective was to address the interoperability between evolving malaria data sources and provide advanced query options [34] for users with little or no technical skills. The platform leverages Semantic Automated Discovery and Integration (SADI) [31] Semantic Web Services and a semantic query engine HYDRA [32] to implement the target queries typical of malaria programs. The platform uses community-developed Malaria ontologies, to describe data services. It enhances the findability of distributed data resources, and the construction of workflows to fetch data from different Web services.

Al-Manir *et al.* [28] reported on use cases provided by the Uganda Ministry of Health to illustrate effectiveness in providing seamless access to distributed data and preservation of interoperability between online resources. Specifically, the queries investigate the nature of interventions *e.g.* which indoor residual spraying used permethrin as an insecticide?, and more complex queries looking at the impacts of interventions *e.g.* which districts of Uganda that used permethrin-based long-lasting insecticide-treated nets in 2015 saw a decrease in *Anopheles gambiae* s.s. population but no decrease in new malaria cases between 2015 and 2016? This latter query is a particularly complex query that involves a combination of multitude of services discovered and orchestrated into a workflow by the HYDRA query engine (See Figure 1 for the list of registered services).

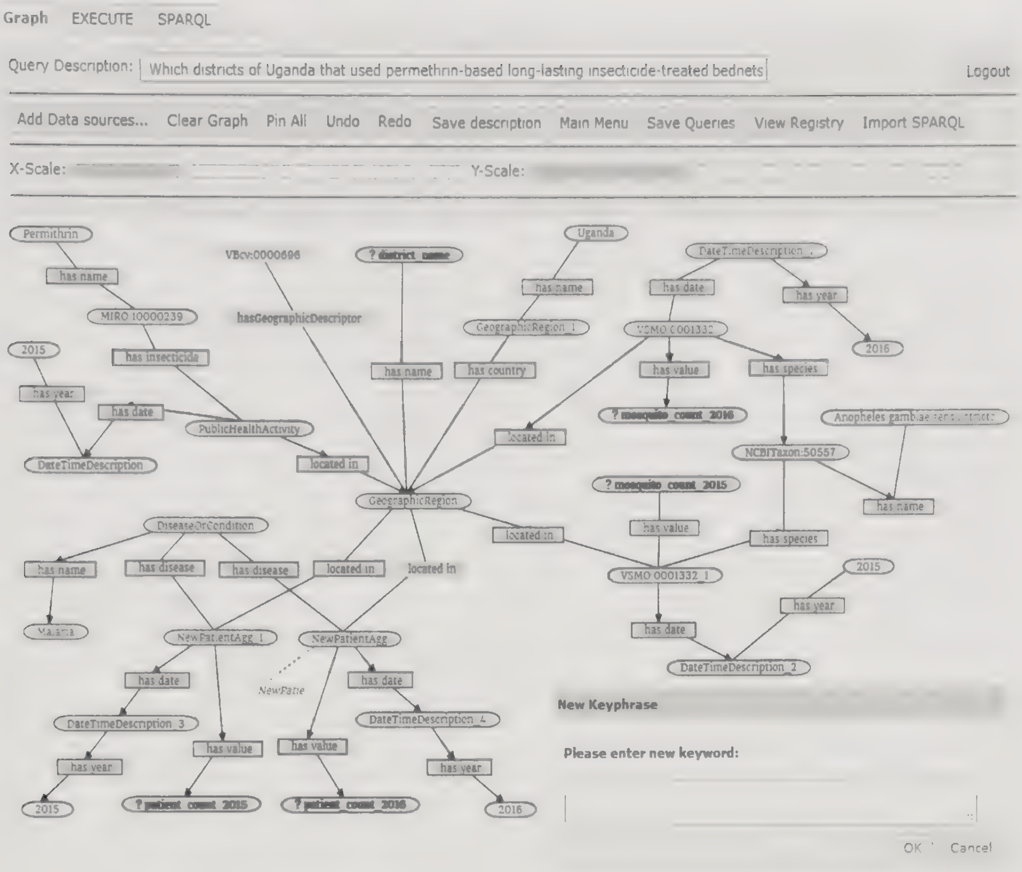


getSpeciesIdByPopulationId
allMosquitoPopulations
getSpeciesIdentificationMethodDescriptionByPopulation
allFieldPopulations
allAssays
allCollectionSites
getCollectionSiteIdByPopulationId
getCountryByCollectionSiteId
getInsecticideIdByAssayId
getPopulationIdByAssayId
getResultByAssay
getNameByGeographicRegionId
getNameByHouseholdId
getHouseholdIdByPublicHealthActivityId
getGeographicRegionIdByHouseholdId
getDateByPublicHealthActivityId
allPublicHealthActivities
getNameByPublicHealthActivityId
getGeographicRegionIdByPublicHealthActivityId
getCountByThing
getInsecticideIdByIndoorResidualSprayingId
getYearByDate
isGeographicRegionIdADistrict
getInsecticideIdByLongLastingInsecticidalNetId
getNameByInsecticideId
getCountryIdByGeographicRegionId
getGeographicRegionIdByEstimationOfSizeOfOpenInsectPopulationId
getDateByEstimationOfSizeOfOpenInsectPopulationId
getValueByEstimationOfSizeOfOpenInsectPopulationId
getSpeciesIdByEstimationOfSizeOfOpenInsectPopulationId
getNameBySpeciesId
getGeographicRegionIdByNewPatientAggId
getDiseaseIdByNewPatientAggId
getNameByDiseaseId
getDateByNewPatientAggId
getValueByNewPatientAggId

Figure 1. List of registered services.

Figure 2 shows a graphical query presented in the HYDRA GUI and the corresponding SPARQL query. Keyword /graphical inputs presented on a canvas are converted to SPARQL queries and presented to HYDRA for processing. These semantic queries are sharable, editable, and offer a high degree of intelligibility for surveillance experts. There is significant flexibility provided to compose regular and ad hoc queries independent of users needing to understand a data structure or a query syntax. Queries posed to the SIEMA surveillance platform are translated into workflows of services by HYDRA which are composed of one or more SADI services identified in the registry.

For explainability the service interface of each SADI service is based on the myGrid/Moby service Ontology [35] which requires that a service contains information such as its unique name, the URI to locate the service, the URIs where the input and output are defined, and a textual description. Information about the input and output of the service can be explained from the concepts and relations used in their definition. The concepts and relations, which are derived from community adopted standard vocabularies, are all resolvable through their URIs. The services themselves are resolvable from their URIs. Thus, the workflow is explainable as each service and the associated metadata about the input and output of a service is explainable.



Query Description: Which districts of Uganda that used permethrin-based long-lasting insecticide-treated bednets

New Keyphrase

Please enter new keyword:

```
PREFIX ex: <http://example.org/>
SELECT  ?district_name ?mosquito_count_2015 ?mosquito_count_2016
?patient_count_2015 ?patient_count_2016
WHERE
  { ?NewPatientAgg          a  ex:NewPatientAgg .
    ?GeographicRegion       a  ex:GeographicRegion .
    ?VBcv:0000696           a  ex:VBcv:0000696 .
    ?GeographicRegion       a  ex:GeographicRegion .
    ?GeographicRegion_1     a  ex:GeographicRegion .
    ?PublicHealthActivity     a  ex:PublicHealthActivity .
    ?DateTimeDescription    a  ex:DateTimeDescription .
    ?MIRO:10000239          a  ex:MIRO:10000239 .
    ?NewPatientAgg             ex:located_in  ?GeographicRegion .
    ?GeographicRegion          ex:hasGeographicDescriptor  ?VBcv:0000696
.
    ?GeographicRegion          ex:has_country  ?GeographicRegion_1 .
    ?GeographicRegion_1        ex:has_name  "Uganda"^^xsd:string .
    ?GeographicRegion          ex:has_name  ?district_name .
    ?PublicHealthActivity      ex:located_in  ?GeographicRegion .
    ?PublicHealthActivity      ex:has_insecticide  ?MIRO:10000239 .
    ?MIRO:10000239             ex:has_name  "Permithrin"^^xsd:string .
    ?PublicHealthActivity      ex:has_date  ?DateTimeDescription .
    ?DateTimeDescription    a  ex:DateTimeDescription ;
                               ex:has_year  2015 .
    ?VSMO:0001332           a  ex:VSMO:0001332 .
    ?VSMO:0001332              ex:located_in  ?GeographicRegion ;
                               ex:has_value  ?mosquito_count_2016 .
    ?DateTimeDescription_1 a  ex:DateTimeDescription .
    ?VSMO:0001332              ex:has_date  ?DateTimeDescription_1 .
    ?DateTimeDescription_1     ex:has_year  2016 .
    ?NCBITaxon:50557        a  ex:NCBITaxon:50557 .
    ?VSMO:0001332              ex:has_species  ?NCBITaxon:50557 .
    ?NCBITaxon:50557           ex:has_name  "Anopheles gambiae  sensu
                               stricto"^^xsd:string .
    ?VSMO:0001332_1         a  ex:VSMO:0001332 .
    ?VSMO:0001332_1            ex:has_species  ?NCBITaxon:50557 .
    ?VSMO:0001332_1            ex:located_in  ?GeographicRegion .
    ?DateTimeDescription_2 a  ex:DateTimeDescription .
    ?VSMO:0001332_1            ex:has_date  ?DateTimeDescription_2 .
    ?DateTimeDescription_2     ex:has_year  2015 .
    ?VSMO:0001332_1            ex:has_value  ?mosquito_count_2015 .
    ?NewPatientAgg_1        a  ex:NewPatientAgg .
    ?NewPatientAgg_1           ex:located_in  ?GeographicRegion .
    ?DiseaseOrCondition     a  ex:DiseaseOrCondition .
    ?NewPatientAgg_1           ex:has_disease  ?DiseaseOrCondition .
    ?DiseaseOrCondition        ex:has_name  "Malaria"^^xsd:string .
    ?NewPatientAgg             ex:has_disease  ?DiseaseOrCondition .
    ?DateTimeDescription_3 a  ex:DateTimeDescription .
    ?NewPatientAgg_1           ex:has_date  ?DateTimeDescription_3 .
    ?DateTimeDescription_3     ex:has_year  2015 .
    ?NewPatientAgg_1           ex:has_value  ?patient_count_2015 .
    ?DateTimeDescription_4 a  ex:DateTimeDescription .
    ?NewPatientAgg             ex:has_date  ?DateTimeDescription_4 .
    ?DateTimeDescription_4 a  ex:DateTimeDescription ;
                               ex:has_year  2016 .
    ?NewPatientAgg             ex:has_value  ?patient_count_2016 .
  }
FILTER (?mosquito_count_2015 < ?mosquito_count_2016)
FILTER (?patient_count_2015 >= ?patient_count_2016)
```

Figure 2. SPARQL query and graphical representation for "Which districts of Uganda that used permethrin-based long-lasting insecticide-treated nets in 2015 saw a decrease in *Anopheles gambiae* s.s. population but no decrease in new malaria cases between 2015 and 2016?"

The system described in [28, 33] serves as a functioning prototype funded by the Bill and Melinda Gates Foundation Grand Challenges. It demonstrates the state of the art with respect to surveillance. Further deployment of this approach will likely emerge for other surveillance tasks given that the terminology layer and registry used in the framework can be exchanged for other specific terminologies.

Maintaining Interoperability

Another challenge that is symptomatic of existing surveillance systems is that they can be brittle, in the sense that even minor updates to core terminologies by domain experts, which occur regularly and incrementally, can render them inactive. The challenge is to preserve the integrity and consistency of an integrated system (interoperability). Here the primary activities during this change management activity are detection, representation, validation, traceability, and rollback, as well as the reproduction of the changes [36]. Mitigating this challenge is also a feature of the SIEMA platform [33] which incorporates a custom algorithm to detect changes to community-developed terminologies, data sources, and services and reports these details to a dashboard displaying real-time service availability (uptime/downtime). Based on the type of change detected and its impact on the status of a service, the dashboard is updated. This level of reporting makes it possible for surveillance practitioners to invoke Valet SADI [37] to automatically rebuild services as needed, mitigating service downtime to ensure reliability.

Conclusion and Outlook

It is clear that pandemic preparedness is an essential theme for the future. Smart surveillance centers and observatories will be required for each municipality and regional governments to archive a range of key datasets. Global Health Observatories [38] have already recorded more than 1000 key indicators for 194 WHO's member states. Smart cities will need to incorporate provisions for effective interventions, that will need to target both commercial and residential sectors, such as social distancing measures required during pandemics. Primary data, including sensor datasets, need to be readily available for secondary uses in unanticipated ways to improve decision making by governments and civic leaders while ensuring privacy protection and adhering to ethical standards and guidelines

The integration of such datasets will be of paramount importance and modeling of these data for reporting purposes must be prepared in advance. Semantics and ontologies will play an essential role in ensuring interoperability and rapid reuse of data sets for reporting. Preparedness as an activity must include dynamic data sources and all aspects of digital infrastructures that support decision making. In the context of the 2020 Covid-19 pandemic, the Ontology of Coronavirus Infectious Disease (CIDO) [40] was developed to provide standardized human- and computer-interpretable annotation and representation of various coronavirus infectious diseases, including their etiology, transmission, pathogenesis, diagnosis, prevention, and treatment. More specifically the COVID-19 Surveillance Ontology [41] is an application ontology used to support COVID-19 surveillance in primary care. The ontology facilitates monitoring of COVID-19 cases and related respiratory conditions using data from multiple brands of computerized medical record systems. It is anticipated that these ontologies will be expanded by integrating further knowledge from relevant domains to provide a comprehensive semantic backbone necessary for intelligent global pandemic surveillance and policymaking that is essential today.

References

[1] Zheng J, Cade J, Brunk B, Roos D, Stoeckert C, Sullivan S, et al. Malaria study data integration and information retrieval based on OBO Foundry ontologies. In: CEUR Work-shop Proceedings 1747. 2016 Presented at International Conference on Biological Ontologies; Aug 1-4, 2016; Corvallis, Oregon, USA p. 1-4.

[2] Mboera L, Rumisha S, Mlacha T, Mayala B, Bwana V, Shayo E. Malaria surveillance and use of evidence in planning and decision making in Kilosa district, Tanzania. Tanzan J Health Res 2017;19(3):1-10

[3] Ohiri K, Ukoha NK, Nwangwu CW, Chima CC, Ogundeji YK, Rone A, et al. An assessment of data availability, quality, and use in malaria program decision making in Nigeria. Health Syst Reform 2016 Sep 23;2(4):319-330.

[4] Ohrt C, Roberts KW, Sturrock HJW, Wegbreit J, Lee BY, Gosling RD. Information systems to support surveillance for malaria elimination. Am J Trop Med Hyg 2015 Jul;93(1):145-152.

[5] Fu C, Lopes S, Mellor S, Aryal S, Sovannaroth S, Roca-Feltrer A. Experiences from developing and upgrading a web-based surveillance system for malaria elimination in Cambodia. JMIR Public Health Surveill 2017 Jun 14;3(2):e30.

[6] Lu G, Liu Y, Beiersmann C, Feng Y, Cao J, Müller O. Challenges in and lessons learned during the implementation of the 1-3-7 malaria surveillance and response strategy in China: a qualitative study. Infect Dis Poverty 2016 Oct 05;5(1):94.

[7] Ibrahim B, Abubakar A, Bajoga U, Nguku P. Evaluation of the malaria surveillance system in Kaduna state, Nigeria 2016. Online J Public Health Inform 2017 May 02;9(1):e177.

[8] Briand D, Roux E, Desconnets JC, Gervet C, Barcellos C. From global action against malaria to local issues: state of the art and perspectives of web platforms dealing with malaria information. Malar J 2018 Mar 21;17(1):122.

[9] World Health Organization. Global technical strategy for malaria 2016-2030. Geneva, Switzerland: WHO; 2015. http://apps.who.int/iris/bitstream/handle/10665/176712/9789241564991_eng.pdf

[10] The FAIR Data Principle. La Jolla, CA: FORCE11; 2017.https://www.force11.org/group/fairgroup/fairprinciples

[11] Ontologies: vocabularies. Cambridge, MA: W3C; 2015. https://www.w3.org/standards/semanticweb/ontology

[13] Tan, Y. S. (2017). Reconstructing Data Provenance from Log Files (Thesis, Doctor of Philosophy (PhD)). The University of Waikato, Hamilton, New Zealand. Retrieved from https://hdl.handle.net/10289/11388

[14] Godfrey MW, German DM, Davies J, and Hindle A. Determining the provenance of software artifacts. In Proceedings of the 5th International Workshop on Software Clones (IWSC '11). Association for Computing Machinery, New York, NY, USA, 2011:65–66.

[15] Cuevas-Vicenttín, V., Ludäscher, B., Missier, P., Belhajjame, K., Chirigati, F., Wei, Y., & Leinfelder, B. (2015). ProvONE: A PROV extension data model for scientific workflow provenance. Draft 01 May 2016.

[16] Davidson, S. B., Boulakia, S. C., Eyal, A., Ludäscher, B., McPhillips, T. M., Bowers, S., Anand, M. K. & Freire, J. (2007). Provenance in Scientific Workflow Systems. *IEEE Data Eng. Bull.*, 30, 44-50.

[17] Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S. & Zhao, J. (2013). PROV-O: The PROV Ontology. *W3C Recommendation 30 April 2013*. https://www.w3.org/TR/prov-o/

[18] Lozano-Fuentes S, Bandyopadhyay A, Cowell LG, Goldfain A, Eisen L. Ontology for vector surveillance and management. J Med Entomol 2013 Jan;50(1):1-14.

[19] Topalis P, Mitraka E, Dritsou V, Dialynas E, Louis C. IDOMAL: the malaria ontology revisited. J Biomed Semantics 2013, Sep 13;4(1):16.

[20] Dialynas E, Topalis P, Vontas J, Louis C. MIRO, and IRbase: IT tools for the epidemiological monitoring of insecticide resistance in mosquito disease vectors. PLoS Negl Trop Dis 2009 Jun 23;3(6):e465.

[21] Dórea FC, Vial F, Hammar K, et al. Drivers for the development of an Animal Health Surveillance Ontology (AHSO). Prev Vet Med. 2019;166:39-48. doi:10.1016/j.prevetmed.2019.03.002.

[22] Schriml L, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012 Jan;40(Database issue): D940-6.

[23] `Horridge M., Parsia B., Sattler U. (2009) Explaining Inconsistencies in OWL Ontologies. In: Godo L., Pugliese A. (eds) Scalable Uncertainty Management. SUM 2009. Lecture Notes in Computer Science, vol 5785. Springer, Berlin, Heidelberg.

[24] Low, H., Baker, C., Garcia, A. et al. An OWL-DL Ontology for Classification of Lipids. Nat Prec (2009). https://doi.org/10.1038/npre.2009.3542.1

[25] Chepelev LL, Riazanov A, Kouznetsov A, Low HS, Dumontier M, Baker CJ. Prototype semantic infrastructure for automated small molecule classification and annotation in lipidomics. BMC Bioinformatics. 2011;12:303.

[26] Krötzsch M. Ontologies for Knowledge Graphs? Proceedings of the 30th International Workshop on Description Logics, Montpellier, France, July 18-21, 2017. CEUR Workshop Proceedings 1879, CEUR-WS.org 2017.

[27] Xiao G, Calvanese D, Kontchakov R, Lembo D, Poggi A, Rosati R, and Zakharyaschev M. (2018). Ontology-based data access: A survey. IJCAI 2018: 5511-5519.

[28] Al Manir MS, Brenas JH, Baker CJ, Shaban-Nejad A. A Surveillance Infrastructure for MalariaAnalytics: Provisioning Data Access and Preservation of Interoperability. JMIR Public Health Surveill.2018;4(2): e10218.

[29] Dumontier M, Baker CJO, Baran J, Callahan A, Chepelev L, Cruz-Toledo J, Del Rio NR, Duck G, Furlong LI, Keath N, Klassen D, McCusker JP, Queralt-Rosinach N, Samwald M, Villanueva-Rosales N, Wilkinson MD, and Hoehndorf R. The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. J Biomed Semantics. 2014 Mar 6;5(1):14.

[30] van den Driessche P. Reproduction numbers of infectious disease models. Infect Dis Model. 2017 Jun 29;2(3):288-303.

[31] Wilkinson MD, Vandervalk B, McCarthy L. The semantic automated discovery and integration (SADI) web service design-pattern, API and reference implementation. J Biomed Semantics 2011 Oct 24;2(1):8.

[32] HYDRA. Saint John, NB: IPSNP Computing Inc http://ipsnp.com/HYDRA/

[33] Brenas JH, Al-Manir MS, Baker CJO, Shaban-Nejad A. A malaria analytics framework to support evolution and interoperability of global health surveillance systems. IEEE Access 2017;5:21605-21619.

[34] Brenas JH, Al Manir MS, Zinszer K, Baker CJO, Shaban-Nejad A. Exploring semantic data federation to enable malaria surveillance queries. Stud Health Technol Inform 2018;247:6-10.

[35] The myGrid Moby service Ontology. http://www.mygrid.org.uk/mygrid-moby-service

[36] Shaban-Nejad A, Haarslev V. Managing changes in distributed biomedical ontologies using hierarchical distributed graph transformation. Int J Data Min Bioinform 2015;11(1):53-83.

[37] Al Manir, M. S., Riazanov, A., Boley, H., Klein, A., & Baker, C. J. (2016, July). Valet SADI: provisioning SADI web services for semantic querying of relational databases. In Proceedings of the 20th international database engineering & applications symposium (pp. 248-255).

[38] The World Health Organization (WHO) Global Health Observatory data repository: https://apps.who.int/gho/data/view.main

[39] Baker CJO, Warren RH, Haarslev V, and Butler G. The Ecology of Ontologies in the Public Domain, The Monist, Oct 2007, 90(4): 585–601.

[40] He, Y., Yu, H., Ong, E. et al. CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Sci Data 7, 181 (2020). https://doi.org/10.1038/s41597-020-0523-6.

[41] COVID-19 Surveillance Ontology https://bioportal.bioontology.org/ontologies/COVID19

BIOS

Christopher J. O. Baker, PhD.: is Professor and Chair in Computer Science at the University of New Brunswick in Saint John. He has 30 years of expertise spanning in microbiology, biomedical informatics, data integration, interoperability and biosurveillance. Since 2018 he has served on the advisory board of the Canadian Institute for Cybersecurity.

Mohammad Sadnan Al Manir, PhD.: is a postdoctoral researcher at the University of Virginia. His current research focuses on a cloud interoperability framework for FAIR, citable, reproducible sharing of analyses, study results, and their sources. His areas of research include data federation in surveillance, NLP, occupational health and medicine, and semantic technologies.

Jon Hael Brenas, PhD.: is a postdoctoral researcher at the Big Data Institute of the University of Oxford. His expertise lies in formal logics and graph transformation applied to biomedicine and in particular genomics and malaria surveillance.

Kate Zinszer, PhD.: is an Assistant Professor at School of Public Health, University of Montréal and Researcher at the Centre for Public Health Research. She is an expert in infectious disease epidemiology, surveillance, and intervention evaluation for emerging infectious diseases.

Arash Shaban-Nejad, PhD, MPH: is an Assistant Professor in the UTHSC-OAK-Ridge National Lab (ORNL) Center for Biomedical Informatics, and the Department of Pediatrics at the University of Tennessee Health Science Center (UTHSC). He is expert in Artificial Intelligence, Knowledge Representation, Semantic Web and Ontologies, and clinical and public health surveillance. He is the principal investigator in a global health and development research project for malaria elimination, funded by Bill & Melinda Gates Foundation.

Financial Industry Explanations

Mike Bennett

Hypercube Limited, London, UK

Abstract

Accountability is an important requirement of the financial industry. Reporting and explaining can be the means by which accountability is achieved but only if the parties have shared meaning for the terms being used. What makes this difficult is that financial terminology frequently uses simple everyday terms in highly technical ways that differ from one context to another. This article examines the issues for ensuring that financial explanations are correctly understood by all parties, both at micro- and macro-economic levels. The proposed technique for solving this problem is to use ontologies. Several examples of successes with the use of ontologies and business rules in an ontological framework are presented in some detail. Since ontologies are a relatively new technique for the financial industry, it is necessary for the ontology itself to be explainable. This problem is also discussed. The conclusion is that the financial system could benefit from formal approaches to explainability based on ontologies.

1. Introduction

THE FINANCIAL INDUSTRY has a lot of explaining to do.

There are a lot of ways of looking at a financial institution such as a bank. It is an entity that deals with money. It is a data firm with complex data interactions. It is an entity that buys and sells risk.

These things all have a role in explaining. Banks may need to explain to customers why they didn't approve this loan or that line of credit. They need to explain to regulators and shareholders why they did. Regulators look both for microprudential and macroprudential risk – that is, what risks banks take on for themselves and what risks they present to the broader economy.

Explanations form one kind of a more general matter which is central to finance: accountability. Institutions have to account for themselves to their shareholders and to regulators and central banks. Regulators and central banks give an account of things to lawmakers and the public, and so on. The more specific notion of explanation may come into play at any point in this information lifecycle – the key thing is that the relevant data are available, timely, accurate and understandable. Central banks will apply a number of statistical analyses to the data that come in from individual banks in the economy for which they have oversight.

In this article we begin in Section 2 by discussing the relationship among explanations, reports and accountability. This will provide some motivation for why financial explanations are important, both for the financial services industry in general and for the retail finance sector in particular. For explanations and reports to be meaningful, the terms that are used to express the explanations and reports must be understood by all parties that are involved. In other words the parties must have commonly accepted shared meanings for the terms. This is more difficult to achieve than one would expect, and in Section 3, we explain why simple solutions such as glossaries and dictionaries are inadequate and introduce ontologies as a better solution. We then give a variety of examples of successes with the use of financial ontologies in Section 4. While some of the examples are still at the proof of concept stage, the results show great potential for solving difficult problems in financial services. Sections 5 and 6 discuss some of the challenges with the use of financial ontologies. Section 5 discusses the notion of a business rule and the difficulties involved in formulating and enforcing them in an ontological framework. While introducing ontologies to finance has the potential for solving significant problem for explainability, reporting, regulatory enforcement and so on, it adds another problem; namely, explaining the ontology itself. This challenge is discussed in Section 6. We end with some final observations and conclusions in Section 7.

2. Reporting and Accountability

Reporting itself covers a wealth of requirements. Broadly speaking the reporting requirements in finance fall under one of two basic motivations: ensuring that consumers and other participants in the financial system are fairly treated; and making sure that the system itself does not fall into some unstable state.

For risk there is both internal and external reporting, risk management assessment, and compliance. On the consumer protection side there are regulations setting out a number of reporting requirements to ensure fairness towards investors, including price transparency and fairness, compliance with investment guidelines, compliance with stated fund management objectives, and so on.

Much of the focus of domestic regulation before the 2008 Global Financial Crisis (GFC) could be assumed to have been driven by lawmakers, who are driven by their electorates. The GFC provided a wake-up call in that what was needed was neither more regulation nor less regulation, but different kinds of regulation, embodying fresh thinking on the nature of global systemic risk. The sum of regulations that aim to protect the consumer would not sum up to better protection of the financial system itself.

2.1 Macroprudential Risk

The global financial system can be viewed as a complex dynamic system. This means that there are emergent behaviors arising out of actions and interactions within the system. It is not realistic to sum up all the risks seen by each participant in the system and expect to understand the risks to the system as a whole. This became very apparent in the 2008 GFC. For this reason financial system regulatory bodies look for a number of different kinds of information from industry participants, ranging from consumer-level protections (avoiding mis-selling and the like) to macro-economic and systemic risk factors. However, part of the nature of emergent phenomena in complex adaptive systems is that what emerges can't always be explained – at best, we can know why we don't know what happens next. Financial systemic risk management is more about anticipating what risks might be starting to emerge than about accounting for what might have happened after the fact. For this reason there are also initiatives in macroprudential regulation such as the Basel Committee for Banking Supervision BCBS239 regulation (Bank for International Settlements, 2013). This regulation defines a category of 'Systemically Important Financial Institution' (SIFI) and sets out how SIFIs are to be able to submit reports in the future under conditions that are not known in the present. As one central banker put it, we can't expect to deal with the next financial crisis using the reports that were appropriate for the previous one.

The reports and information submitted to regulators are not explanations in the form defined elsewhere in this issue, but provide the raw material for providing such explanations. A further lesson from the GFC was that data on their own are only part of the requirements for understanding what was going on. Many firms had all the data they needed to understand their exposures to failing or at-risk institutions, but it still took several weeks to arrive at a knowledge of those exposures.

Data does not mean knowledge. For that you need the addition of some kind of meaning; the semantics of the data enables them to be understood and re-used as a source of knowledge. For this reason the financial industry, like many others, is starting to figure out how to introduce formal ontologies into these data workflows.

Data accompanied by formal semantics do not in themselves form a set of explanations but they do provide the raw material for explainability. With this in mind we can look at a number of specific examples of explanations in the financial services space. Some of these are made available directly to the user, while others are directed to public regulatory authorities or to other financial ecosystem participants.

2.2 Explanations in Retail Finance

The common source of explanations requirements in retail is, of course, the customer. Whether in retail banking or credit cards, customers will want to know why they have been denied a new card or an increase in their credit. The standard set of answers include specific things like ‘your income is too low’ or ‘your existing credit balances are too high’ but may also include question-begging responses such as ‘You did not meet our criteria’ (what criteria?), ‘Too many credit enquiries’ (how many is too many?) or ‘You have not been at your current job long enough’ (how long is long enough?).

Many of the seeming explanations given to the retail customer will themselves beg follow-up questions that that customer feels they need to know. Set against this, the retail institution is often reluctant to hand over all the models and model inputs that they would use to make these decisions (thereby furnishing a full explanation) for the entirely understandable reason that someone in possession of all of these model parameters may use that information to game the system. Explainability leads to vulnerability.

At the same time, customers have the right to know that decisions made about them are made fairly and equitably. They have a right to know that the decisions made were based on current, up to date and accurate data about them. What if the job they are in is not the one on which the credit decision was based, or the debt balances ascribed to them have recently been paid off?

The challenge for the retail bank or credit institution is to ensure that the data they are working from are complete and coherent. Information they hold needs to be coordinated with that held by third parties such as credit reference agencies, and *vice versa*. Meanwhile their holding of data about each customer must comply with applicable regulations for disclosure on the one hand, such as 'Know Your Customer' (KYC), and non-disclosure on the other, such as the General Data Protection Regulations (GDPR) in the European Union.

Temporal mismatches need to be avoided, for example the lag time between data reported on at defined intervals and the data about things as they stand at the present time. Then there are variations in the usual pattern of credit card or *current* account holding, such as accounts with multiple holders, or those with holders who come under a status with specific protections, such as veterans. In some jurisdictions information held against an address, such as county court judgments, often causes subsequent occupants of that address to get down-graded – usually without a clear explanation that this is the case. Common complications often arise from situations of the borrower such as divorces, court orders, changes of address, or other personal circumstances. On the happier side of things there are unscheduled pre-payments by the borrower, where this is allowed. There may also be situations not under the control or knowledge of the customer such as concurrent fraud investigations, automatic payment system delays, differences or misunderstandings on month-end roll-over dates and so on.

This complex set of interlocking processes, disclosure and non-disclosure requirements and mismatches in knowledge between one party and another add up to a complex data management problem. It is also a complex problem of managing the relationships between the data and the things in the real world that the data are about.

The best-known of these is the issue of 'bi-temporality' in data management. This is the distinction between the date or time for which the data are about is available to the decision-maker, and the date or time that something occurred in reality for which the data are about. This matter of time is just one respect in which the state of things in reality and the state of the data that aim to reflect this may differ. Processes for data management and information supply chains need careful design and management, taking into account who needs to know what, how often and in what level of detail something needs to be reported, and how changes in circumstances are propagated across the entirety of the systems that hold relevant information, even understanding the impact on different data resources of a specific kind of change in the underlying reality.

Good customer service requires understanding the customer's journey through life. This sounds a bit like some set of buzz-words, but in reality it is important for the institution to be able to follow and understand the customer's changing situation, along with changes in statute law and in the institution's internal lending practices and longer term strategies, simply in order to avoid unnecessary misunderstandings, unhappy customers, reputational risk or legal exposures.

Explaining things in retail then is harder than it seems. Call center employees, the usual point of contact between the institution and the borrower, are effectively playing the role of knowledge workers, but all too often the knowledge they need is not available, or the knowledge is available but they are not skilled to the required level to make use of it in initial customer contacts. The tools they use may not be interoperable across different data sources, or the data they need may not be available in real time.

Meanwhile explanations are by their nature very contextual or scenario-dependent, and these dependencies also need to be understood. Decision support software meanwhile needs to balance the requirements for customer retention, profitability, and regulatory compliance.

These challenges will only get greater as innovations arise, both in technology, such as the increasing use of artificial intelligence in decision making, and in the financial marketplace itself, both in the emergence of new financial models in micro-finance and in the emergence of new technology-based systems such as distributed ledger (blockchain) technologies.

3. Shared Meaning

In the move towards more coherent use of data in financial reporting and decision making, one common theme has been the requirement for some way to represent common, shared business meanings. The common reaction to this requirement has been to try to establish business dictionaries or glossaries, where terms (words) are standardized to always mean the same thing, so this can be used as a point of reference across different data resources (Knight, 2018). A similar approach from the technology side is the use of 'data dictionaries'.

Both of these approaches suffer from a common weakness. In data dictionaries, each data model has textual information giving a 'definition' against each data element (field names and the like). It does not take long for these definitions to start to sprout extra qualifications, of the form "in the case of (XYZ) instrument, this field represents (some specific thing)." Soon extra business rules are added, reflecting logical statements about what sort of thing should be in a given field under different circumstances. The reason for this is that in any good data model design, data elements do not map precisely to single meanings of things in the real world. If they did, they would not be a design; there would be no data normalization or re-use.

Business dictionaries or glossaries have the same weakness but for a different reason. The way that humans use words is very contextual. I can use a word like 'bank' and you will know what I mean by whether I am talking about investment or fishing. The same words mean different things in different contexts. Even within finance itself there are subtle differences, for example the term 'over the counter' might refer to a derivative trade that is struck directly between parties, or it might refer to securities that are traded directly rather than through an exchange. A subtle difference but again any dictionary would need to add qualifying terms to state what concept is referred to by the words in different contexts. Meanwhile different parts of the industry and different functions within a firm may use different words to refer to the same concepts, for example 'coupon' or 'interest' on a debt instrument.

With both data dictionaries and business dictionaries this is not a fault but a feature – neither human words nor data field names map directly to concepts. The push for 'Why can't we all agree on the same terms' inevitably comes up against what Wittgenstein in his later work (Wittgenstein, 1953) calls 'language games'. Words play games.

People in the financial industry have long sought to solve the question of shared meaning for data elements, for example under the guidance of the Enterprise Data Management Council (EDM Council, n.d.). It was during one such meeting in which people were trying to agree on common terms for 'Critical Data Elements' in securities clearing and settlement, that someone thought to ask the question: "While we are disagreeing on what words to use, do we at least agree about what the concepts are?" The answer was yes – everyone agreed what the real world meanings, the concepts, were.

From this realization, the idea for a common ontology for the financial industry was born. The Financial Industry Business Ontology (FIBO) (Bennett, 2010) was initially conceived as a source of common shared meaning for the industry, to provide a point of reference for data models, integration, reporting, and other requirements. This was subsequently standardized as a series of machine-readable ontologies for use with financial industry data in a range of applications.

A formal ontology provides a simple account of the meanings of things. It does this by means of declarative statements of the form 'there exists' and qualifications such as 'for all'. This falls under what is defined as first-order logic, and most published ontologies for use with data (including FIBO) use a sub-set or variation on FOL called Description Logic (DL). This is the sub-set of logic for which it is possible mathematically to prove that the assertions in the ontology are consistent and can be reasoned over in a finite period of time.

An ontology simply sets out a logical definition of what kinds of things there are, and what features distinguish one thing from another. FIBO defines a range of financial instruments in these terms, along with kinds of business entities and the relationships between these.

This way of saying things about something in the real world is necessarily limited but useful; in plain English terms, this first-order kind of ontology defines what there is, what kind of a thing something is, and what features or characteristics of a thing distinguish it from other things; that is, what are the necessary and sufficient characteristics for something to be considered to be the member of a particular set of things. This set-theoretic notion effectively defines a 'concept' (Odell, 2011). More specifically this is an 'intentional' definition of a concept. Some ontology languages also allow for extensional definitions where a set of things is defined by explicitly specifying all of its members.

4. Financial Ontology Examples

Some uses of ontology rely on the technical deployment of these as part of a solution to a specific problem, for example to draw inferences from available data. Other uses, in data management, integration, and reporting as well as artificial intelligence, rely on the provision of common meaning, via formally defined concepts, to streamline the information supply chain for reporting, for example, clarifying the meaning of each item in a report. This aids in the accountability of data in financial regulatory reports and thereby the explainability of the information contained.

This basic reflection of reality, as exemplified by FIBO, can be used to gain insights from data that would normally be sitting in different data silos under different schematic structures. For example, one set of data might contain information about a series of derivatives transactions, while another data source would carry information about the ownership and control relations between corporations or other business entities.

The underlying abstract model for many ontologies, the Resource Description Framework (RDF) (World Wide Web Consortium, 2014) coupled with the Web Ontology Language (OWL) (World Wide Web Consortium, 2012) provides a common syntax, so that terms from different data sources are framed using the same underlying technology language. In the case of RDF, this is the language of 'triples' (relations of the form subject-predicate-object). A collection of triples is a graph in which each triple is an edge in the graph.

The OWL language sits on top of RDF, and if the terms from different data sources in RDF are defined with reference to a single OWL ontology or a single mutually coherent set of ontologies, then data comparisons, formal inferences and semantic queries can be made against that data. FIBO provides one such set of mutually consistent ontologies, covering financial instruments, business entities and entity ownership and control relationships.

This kind of graph-based representation of data, coupled with a common schema in the form of an ontology, is known as a 'knowledge graph'. A precise definition of the term 'Knowledge Graph' is a matter of ongoing discussion in the industry, see for example Ontology Summit (2020). For a workable definition see Yu (2020).

The potential for this basic knowledge graph framework is that if existing data across instrument transactions or holdings and business entity ownership and control hierarchies can be ingested from their existing data habitat and reframed as RDF data under a suitable ontology, we can ask new questions such as, "what is this bank's exposure to that other bank, based on the trading positions it has open with not only that bank but its subsidiaries, parents and affiliates".

4.1 Counterparty Exposures Proof of Concept

A proof of concept to demonstrate this usage was initially carried out at Wells Fargo using FIBO with indicative dummy data (Newman & Bennett, 2012). This was later repeated at another major US bank, State Street, with real data on interest rate swap transactions (David, 2016).

In this proof of concept a set of data about swaps transactions in a standard XML messaging format called FpML (International Swaps and Derivatives Association, n.d.) was fed into the triple data store and each data element was linked to the corresponding term in FIBO to define its meaning. A further set of data, available from U.S. Securities and Exchange Commission (SEC) filings and company registry information, was fed in to define the various ownership and control relations across the institutions that were the counterparty to each swap transaction in the swaps transactions data.

The business motivation for this work was what happened in the Global Financial Crisis: what would happen to the positions at this bank, if a certain other bank were to fall into bankruptcy? Given that each of these institutions has quite complex ownership and control hierarchies, this was not simply a matter of what trades this bank has with that other bank, but what trades it has with its parents and subsidiaries and what the knock-on effect would be on just one of those entities going down, on the complex network of relationships and positions it is tied to.

The resulting knowledge graph was queried using semantic queries, to return data about the monetary amount of each instrument position, the relative capitalization of each institution and the relationships between the relevant institutions. These results, in the form of data, were fed into a visualization program to provide a graph in which the relative trade positions were reflected by the thicknesses of lines between the entities, the capitalization was shown as the size of a circle representing each entity, and ownership and control relationships were additionally represented as lines between entities in different formats.

This proof of concept shows what can be done with ontology, not acting on its own but as something from which to feed graphical visualization techniques. A similar framework could also be used to feed mathematical models or other programmatic solutions. This is not ontology working alone but as a means to integrate data across a range of data sources and carry out operations across that data.

This also shows the use of a particular kind of ontology. In this case the ontology reflects the common meanings of instruments, transactions and business entities, but does so in a way that is directly applicable to data itself.

4.2 Bank of England Proof of Concept

At the Bank of England a pilot project was undertaken to show what could be done with ontology in the reporting chain (Bholat, 2016).

In the existing state of affairs the bank sends out a number of forms to those banks that fall under its jurisdiction. Each box in each form asks for a response, for example to give the amount of debt held by the reporting institution in US Dollars with 3 to 5 years residual maturity. The reporting bank looks to its various internal systems to find the answer to that question and puts it in the form.

Two issues were apparent in this approach. One was that of finding the right information in the right system, an inefficient and potentially time-consuming process. The other was that having received these forms from all these banks, the Bank was not fully confident that each reporting entity had assumed the same intended meaning for each entry in the form; they lacked confidence in the ability to compare like with like.

The premise of the proof of concept was that it should be possible to save time and cost for the reporting entities and at the same time increase the central bank's confidence in the reported information, if reports were made using granular, semantically-aligned data.

A further potential benefit was flexibility for the central bank. If data could be reported in a granular way with clear semantics for each element, then they should not need to send out forms at all. Instead each box on the report would be a semantic query against that granular data. This meant that if the bank wanted to introduce a new box on the form – for example if they wanted to isolate US Dollar debt holdings at different maturities (say, everything with up to 2 years residual maturity) then they did not need to redesign a form with this new box and send it around, they merely needed to write a new semantic query internally and apply this to the same data.

The first step in this proof of concept was to select three forms at random, analyze each line entry, and define the meaning of the data in that box. For example, you may have wondered what the term 'residual maturity' means in the above examples. This is the length of time, on any given day, until the debt in question has been paid off. That is not the same as 'original maturity', which is the length of time to maturity (debt repayment) at the time the security was issued. These may both be given the label 'maturity' in different data models, where the context is obvious by the function of that particular model. The original maturity of a 5 year bond will always have been 5 years, while the residual maturity (or current maturity, or some other label) is the amount of time from today until it matures. A five-year bond issued four and a half years ago has a residual maturity of less than one year, and this determines what box it should be reported in for this example.

The result of this phase of the proof of concept work was a formal ontology of the concepts that the bank had in mind when defining the information that they required in each box. Reporting against this ontology would enable the bank to recreate these forms locally using semantic queries. This would use a data querying language designed for this purpose, called SPARQL (pronounced 'sparkle').

4.3 Regulatory Proof of Concept

More complex reporting requirements call for more complex solutions, including the use of formal business rules.

Regulation W is a US Federal Reserve regulation that establishes terms for transactions between banks and their affiliates (U.S. Electronic Code of Federal Regulations (2002). It was enacted by Congress as part of the Federal Reserve Act and applies to all federally-insured depository institutions.

The Reg. W Proof of Concept initiative (Grosof *et al*, 2015) was formed by the Enterprise Data Management Council (EDM Council) and included Wells Fargo Bank, Coherent Knowledge Systems, SRI International, and the Governance Risk and Compliance Technology Centre (GRCTC) of Ireland (Governance Risk and Compliance Technology Centre, n.d.), with participation from other members of the EDM Council. This combination of participants was selected in order to have access to a range of rules-based technology solutions alongside the basic semantics expertise for the use of FIBO.

Regulation W defines a set of limitations against an illicit market practice called 'front-running'. Front running is the practice of buying or selling a security with advance knowledge of pending transactions that could influence the price, in such a way as to capitalize on that knowledge. To explain or account for whether a given trade does or does not count as a front-running trade, banks that could potentially carry out such trades are required to report all potentially applicable trades under a Regulation W reporting requirement.

In addition to the requirement to account for transactions as not falling foul of the Reg. W requirement, affected banks have an obvious need for internal decision support to determine, ahead of carrying out some potential trade, that it would not fall foul of the Reg. W requirements. This is an explanation requirement.

Core concepts used in this regulatory requirement include 'bank affiliate', 'covered transaction' (a potential transaction covered by the regulation), 'collateral requirements' and the notion of 'low quality assets'. Each of these terms needed to be defined and those definitions acted upon in decision making and explanations. These concepts are used to define limits to potential investments by the firm itself. The regulation stipulates that covered transactions with an affiliate cannot exceed 10 percent of a bank's capital stock and surplus, and transactions with all affiliates combined cannot exceed 20 percent of the bank's capital stock and surplus.

The term 'affiliate' here presented some definitional challenges, since it is both broader and narrower than the normally understood concept of 'affiliate' as being some entity that is either a parent or subsidiary of a given entity. It is broader because Reg. W 'Affiliate' includes firms to which the bank gives certain kinds of investment advice, and narrower since it refers not to the affiliates of all kinds of entity, but only to those that are affiliates, in this broader sense, to a bank. This means that the ontology needed to define the Reg. W concept of ‘affiliate’ as a sub-set of the union of affiliation and investment advisement in relation to the bank for whom the calculations are being carried out.

This is a good (if niche) example of why words alone cannot be used as the basis for meaning. It also illustrates how the meanings of words as defined in specific legislation texts are not necessarily suitable as a source of meaning for those words more generally. What a word means in the context of a given regulation may or may not also be what it means in some different context or some broader set of contexts. This is the reason that institutions need to navigate the realm of meaning by means of concepts and not by words.

The kind of trade that would fall foul of the Reg. W anti-front-running regulation thus required some fairly complex logic to describe it. The use of a formal ontology such as FIBO provides part of the solution to this reporting, in terms of common shared meanings, but there need to be more complex, or higher-order, logical operations on the data in order to determine and explain whether or not each trade comes under the Reg. W limitations.

The aim this proof of concept was to unambiguously understand and automatically comply with regulatory rules. The project used the FIBO in combination with advanced semantic rules defined in the rules languages Rulelog (Grosof, 2013) and Flora-2 (Yang *et al*. 2003), to automatically keep a bank in compliance as transactions were being processed. The intended result was to address the question ‘Am I in compliance?’ This had the associated explanatory requirement ‘Why / why not?’

FIBO and Rulelog/Flora-2 were used to make Reg.-W requirements explicit and applied to sample transaction data to automate compliance assessment. Detailed explanations were provided so that humans could understand the reasoning and facts that led to the conclusions. GRCTC provided expertise in controlled natural language for rule authoring via OMG's Semantics of Business Vocabulary and Business Rules language (SBVR) standard (Object Management Group, 2019) using the SBVR form of structured English. Coherent Technology and SRI technology provided automated reasoning capabilities using the Episto and Sunflower languages, with detailed explanations in English. SRI's technology provided automatic import of knowledge graph data in OWL, into the Flora-2 engine.

The rules engines defined a number of types of transaction that are defined as ‘covered transactions under the regulation, and a number of exemptions that were applicable. The proof of concept demonstrated that using these facilities a bank was able to enter into a transaction with a Counterparty, check if the counterparty is an affiliate, check if the transaction type is covered by regulation W and verify if the amount and total amount are permitted.

The structured English in SBVR was used to capture the business domain, specifically terms referring to business concepts, relationships between concepts and definitional constraints on these relationships. The ‘Rules’ part of the standard was used to capture the business behavioral constraints, obligations, prohibitions and so on. This formed a kind of bridge between the concept language of FIBO for the basic instrument concepts and the technology-based rules languages mentioned earlier.

The methodology developed by GRCTC delivered a system that could follow reference chains and produce self-contained sentences; define terms iteratively until all confusions were clarified; identify, describe and constrain links/relationships between terms, and capture regulatory requirements using the interlinked vocabulary elements from these other steps.

This proof of concept demonstrated the ability to deliver improved confidence in the correctness of compliance checks both for banks and for regulators. This was largely because understandable explanations were provided. This can reduce cost and risk due to the ability of this approach to adapt more easily and quickly to changes in regulations, since these are now framed using a common financial language, aligned with industry standards.

4.4 Explanations in Accounting: Tax Filing Example

Another particularly striking example of explainability makes use of knowledge graphs in an innovative way. This is the system developed by Intuit (Yu, 2020), the software vendor behind QuickBooks and other accounting applications for small businesses and consumers. In this patented innovation, users are able to file tax returns and interrogate each line entry to determine how that figure was derived.

This product uses a knowledge-graph (KG) based solution to determine the values to be placed in each line entry in a tax return. The basic

KG structure is enhanced by the addition of arithmetic functions such as 'add' or 'subtract', these functions being included in the graph structure.

Explanations for a given line entry are then provided to the user by means of traversing the graph to identify each of the inputs and functions used. These can be traversed iteratively so that the input to one function is traced to the output of an earlier function. So, for example, if a tax withholding entry is based on the following rule:

> **20.** If Line 19 is more than line 16, subtract line 16 from line 19.
> This is the amount you **overpaid.**

Then the user can interrogate the line entry for line 20 and see the amounts for lines 16 and 19. They can also traverse the graph by interrogating the line entry for line 19 and determine the line items that went into this and the fact that (in this example) these were added. And so on.

4.5 Ontology in Understanding Data

The range and complexity of requirements for financial institutions to be able to provide accountability and explanations leads to a set of complex data management requirements. One use of ontologies is in assisting the data management function within such firms to have a better understanding of their own data; better explanations internally of the data that they hold and the conclusions that are derived from this data.

These internal explanations make use of something called a 'semantic data catalog' (Newman, 2020). This enables the user, in this case, someone in the data management function within the bank, to pose questions such as "What types of customers are in the Customer table?" or "Where can we find organizational names in this database?"

The knowledge graph provides 'chains of meaning' relating to the real-world subject matter, such as 'Customer has identity some Person', 'Person has name some Personal Name' and 'Personal Name has First Name some string'. These chains of connections make up the ontology and this can be applied to the data held in various databases and mapped to these meanings to provide answers to the questions posed by the data owners.

Similarly the data administrator can ask questions about what information is held in a given data resource, for example "What information is held in the Marketing Database Customer table?" or "Where would I find personal contact information in the Marketing database?" The Semantic Data Catalog is organized in such a way that the user can ask a number of broad based questions about the data.

The ability to address such questions of the data relies on semantic search capabilities, that is, the ability to frame questions in a semantic query form. While this is not explanation, being able to return data based on the semantics of a question is a prerequisite to accessing the right data for accountability or explainability further down the line. Turning user requirements for explanations into formal semantic searches will itself make use of a number of techniques. These include predicting concepts from string values or predicting a vector of concept plus predicate. Predicting string values may use lexical predictions, where mistyped text is replaced with text corresponding to the entries in the knowledge base, using metrics such as the Levenshtein Distance (Levenshtein, 1966); or it may use a concept vector approach, for example replacing the search text 'vanilla interest rate swap' with the synonymous term in the ontology, 'fixed float interest rate swap'. Semantic search using concept and predicate combinations would use a standard semantic querying language directly, for example to return the concept of 'agreement' for a search on 'contract is a type of?'.

This ability to link an ontology to internal data structures also provides the user with metrics on data quality, for example ensuring that stored data conforms to specific patterns for identifiers and the like, or that information on something like a person or a corporation contains all the relevant information as expected for that category of thing, as defined in the ontology. This makes use of another Semantic Web standard called SHACL (World Wide Web Consortium, 2017), which allows the user to define allowable patterns within the data in a knowledge graph.

Given data that may or may not conform to the pattern set out in this pattern (shapes) language, a validation report is produced which flag whether or not the information held in that data source conforms with the requirements for such data – for example that data about a human shall have only one date of birth, a given set of identifiers and so on. Where the report indicates that the data is not conformant to the required pattern an explanation is also generated, showing where the data diverges from the stated requirements.

This ability to validate available data can potentially be applied not only to the management of data, but in managing the requirements for explanations to bank customers, regulators, compliance officers and other end users. For example, a report on some business activity or proposed action, such as investment or lending decisions, may flag up an indication of whether the proposed activity would be conformant to a particular internal rule or external regulation. Simply putting up a flag to say conformance is 'true' or 'false' is not enough; there also needs to be some explanation for this result. These explanations can be derived from the semantic representations of the data in the ontology, as seen in the earlier example for front-running. In practice the various techniques of formal ontology, business rules, semantic queries and semantic 'shapes' can be used in combination to provide explanations to end users, data managers and other stakeholders.

5. Explanations and Logic

When considering the application of rules engines to business compliance and explanations, it is important to define what a business rule really is. It is easy to define technology-based rules engines and apply these to data in some technical ecosystem. It is also easy to claim these are 'business rules'. However, in order for a rule to be a business rule, we should consider the nature of rules: a rule is some piece of logic, applied to something. The logic might be 'if this, then that' or it might be simply 'don't do that'. But what is the 'something' to which the rule is applied? If rules are applied to raw data – that is if the predicates of the rules are data in some system, then there is no guarantee that the rules represent business relationships among business concepts. For this to be the case, the predicates to which the rules are applied must themselves reflect real-world items – that is, concepts in an ontology. The predication of rules determines what kinds of rules they are – application-specific or business rules. Regulatory compliance, accountability and explainability, when they use rules, must use rules that are predicated on some ontology in order to have genuine explanatory power.

The example from Intuit goes beyond the normal usage of a 'first-order' ontology that simply defines what things there are and extends the knowledge graph paradigm, to connect mathematical and logical operations – similar to what we have seen with business rules and semantic shapes, but based in mathematical and arithmetic formalisms (addition, subtraction and so on). Because these features are all aligned with a formal description of 'real world' items (the ontology, assuming a quality ontology that does capture core senses of reality), any good graphical or textual representation of these relationships can be understood by any stakeholder. That is, anyone can derive explanations from a combination of first-order assertions about things in the world, formal rules that operate on those assertions, and mathematical operations on assertions that are of a numerical nature. This is the language of explanation: to the extent that end users can identify what they are seeing with the reality of their world, they should be able to interrogate semantically-enabled data to arrive at their own understanding for the reasons behind data results, decisions and other assertions that are based on that data.

6. Explaining Ontologies

The range and nature of accountability and explainability requirements in finance is indicative of the challenges and opportunities in any data-intensive industry. Formal ontologies of the business concepts are a key component to tying the data to reality and thereby to making data-driven decisions and 'what-if' analyses of proposed actions explainable and furnishing coherent explanations of a financial institution's activities to regulatory authorities.

For this kind of connection to underlying reality to work however, it is important that the ontologies used truly represent the concepts in the business domain. This cannot be approached as 'yet another data modeling' exercise; arguably it should not be approached from within the IT discipline at all. Ontologies need to reflect specialist subject matter. This means that any such ontology needs to be presented to subject matter experts in the relevant business domain, for them to validate and ideally to formally sign off. That means that the content of the ontologies needs to be explained in human-facing ways, whether through tabular views or diagrams. Ontologies are simple declarations of 'What kind of thing is this?' and 'What distinguishes it from other things?' – the necessary and sufficient characteristics for something in the world to belong to a given set of things, even if details of some relationships are harder to formalize. This can be explained relatively simply using set-theoretic notions: set membership, logical unions and intersections and so on.

These basic notions can be presented in a number of possible visual formats – typically simple boxes (or blobs) and lines showing the classes (things) and the relationships between them.

Some more complex logic features used to define the necessary and sufficient conditions are harder to explain. Instead, the ontologist needs to frame specific questions with reference to the classes, for example ‘Is it always the case that one of these has to have this relationship to one of those?’ or ‘Must there necessarily be this property or relationship in order for something to be one of those?’

Unfortunately many of the available tools tend to produce more technology-oriented visualizations, for example auto-generated pictures of ‘bouncy balls’, none of which remain where you left them last time the diagram was generated. This makes it hard to get SME confidence in the basic structure of the model. Some specialist tools do exist that provide a persistent view of the subject matter.

The alternative, which is much to be avoided, is to play into the assumption among many domain experts, that somehow words can be used to solve the problems of meaning. Vocabularies can be generated from an ontology, giving the various words that may be used to reflect a given concept in different contexts. Doing it the other way around – trying to use words as a starting point to represent concepts to the domain experts, is not a good idea. Words play games, and sets of unconnected definitions will give rise to fuzzy, overlapping and incomplete sets of things in the subject matter representation.

The matter of explaining ontologies themselves is currently very immature. Few tools exist, and many ontologies are developed to address specific data-focused application problems (drawing data inferences from existing data, answering specific questions and so on) but are often touted as though they contain some coherent account of meaning. Until these problems are better understood, it is unlikely that we will see the kinds of tools that are needed to ensure that ontologies are adequately explained to business stakeholders, and therefore adequately reflect the business reality that is needed to understand and draw explanations from the wealth of data in large institutions such as in finance.

7. Conclusions

In the world of finance there is much that would benefit from explanation. Formal approaches to explainability are not well established in the industry, but from one perspective the entire financial system can be envisioned as the ebb and flow of data; the raw material of explanations. Some of the requirements for explanations are fairly simple: the reasons for advancing or withholding credit from consumers, for example. Here the industry is moving towards better ways to partner with customers on their life journeys or business trajectories, treating explanations as something to which customers should be entitled. Other explanation requirements are more complex, dealing with the economy, macro-economics, issues of money supply and so on, along with the risk factors that accompany each of these.

Meanwhile the 2008 Global Financial Crisis reinforced an understanding that certain aspects of the global financial system form a complex adaptive (or maladaptive?) system, of the kind from which the phenomenon of 'emergence' gives rise to events and structures that cannot easily be anticipated in advance. Sometimes the best we can do within the parameters of complex systems theory is to explain why we cannot explain something.

In the meantime the massive flows of data in the industry admit of a couple of different purposes – as material for explanations and as something to react to. They provide the source material for accountability, a pre-requisite for explainability. Information reported to regulators can be used to understand and analyze the details that went into why someone or something made a particular decision. Similar data is analyzed by the statistics functions of central banks and used to inform those holding the economic levers of power when and whether to raise or lower interest rates, adjust the money supply and so on. Separately these data flows provide for understanding emergent risks in a system that can never be fully understood, much less explained, but that can be reacted to, given sufficient information.

A common theme in all of these uses of shared data is the need for formal business semantics. In any industry where information technology is extensively used – and many of the issues explored here can be as easily applied in healthcare and elsewhere – the information technology ecosystems used are somewhere in a transition between an older world in which each application had its own data formats and structures, feeding screens or tapes or other things read by humans and not needing to be consumed by different machines, and a future world in which every machine speaks the same language. At this point we have common syntactical formats for exchanging data but not much in the way of common understanding or language.

Confucius was asked what he would do if he was a governor. He said he would "rectify the names" to make words correspond to reality. We now find ourselves in a world where there is more data than there are words to go around, so we need to apply more sophistication to questions of meaning, something that is not an IT function at all but requires deeper business engagement. Even where ‘semantics’ is already being used as a term, it is often in relation to self-contained applications for drawing inferences over limited amounts of data; semantic technology stovepipes replacing rigid database stovepipes, but contributing little to the kind of common language that will be needed to furnish detailed formal explanations of the sort explored in this edition, at every level from consumer protection, through to micro- and macro-economic regulatory oversight and systemic risk mitigation.

References

Bank for International Settlements (January 2013). Principles for effective risk data aggregation and risk reporting. Retrieved September 19 2020, from https://www.bis.org/publ/bcbs239.pdf

Bennett, M. G. (2010). EDM Council Semantics Repository Next Steps – Exploring a Consensus Semantics Framework. In Proceeding: ODiSE'10 Ontology-Driven Software Engineering, Article No. 3 ACM New York, NY, USA 2010. ISBN: 978-1-4503-0548-8. DOI: 10.1145/1937128.1937131

Bholat, D. (2016). Modelling metadata in central banks. ECB Statistics Paper Series. No 13/April 2016. Retrieved 18 September 2020, from https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp13.en.pdf?a66e0b0448ff62df28d39643573b2b3d

David, E., (3 June 2016). FIBO Marches Forward: A Look Inside State Street's FIBO Proof of Concept. Waters Technology. Retrieved 18 September 2020, from https://www.waterstechnology.com/data-management/2459451/fibo-marches-forward-a-look-inside-state-streets-fibo-proof-of-concept

Enterprise Data Management Council (n.d.). Website. Retrieved September 19 2020, from https://edmcouncil.org/default.aspx

Governance Risk and Compliance Technology Centre (n.d.). Website. Retrieved September 19, 2020, from https://www.cubsucc.com/research-centres/governance-risk-and-compliance-technology-centre--grctc-/

Grosof, B. (2013). Rapid Text-Based Authoring of Defeasible Higher-Order Logic Formulas, via Textual Logic and Rulelog. In: Morgenstern, L., Stefaneas, P., Lvy, F., Wyner, A., Paschke, A. (eds.) Theory, Practice, and Applications of Rules on the Web, Lecture Notes in Computer Science, vol. 8035, pp. 2–11. Springer Berlin Heidelberg. Retrieved on 19 September 2020, from http://dx.doi.org/ 10.1007/978-3-642-39617-5_2

Grosof, B., Bloomfield, J., Fodor, P., Kifer, M., Grosof, I., Calejo, M., Swift, T. (2015). Automated Decision Support for Financial Regulatory/Policy Compliance, using Textual Rulelog. Proceedings of the RuleML 2015 Challenge, the Special Track on Rule-based Recommender Systems for the Web of Data, the Special Industry Track and the RuleML 2015 Doctoral Consortium. 9th International Web Rule Symposium (RuleML 2015). Retrieved September 19 2020, from http://ceur-ws.org/Vol-1417/paper8.pdf

International Swaps and Derivatives Association (n.d.). Financial products Markup Language (FpML). Retrieved 19 September 2020, from https://www.fpml.org/the_standard/current/

Knight, M. (January 24, 2018). What is a Business Glossary? Dataversity. Retrieved September 18 2020, from https://www.dataversity.net/what-is-a-business-glossary/

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, Vol. **10**, p.707. Bibcode: 1966SPhD...10..707L

Newman, D., & Bennett, M. (2012). Semantic Solutions for Financial Industry Systemic Risk Analysis. Next Generation Financial Cyberinfrastructure Workshop, Robert H. Smith School of Business, University of Maryland. Retrieved 18 September 2020, from https://wiki.umiacs.umd.edu/clip/ngfci/images/8/80/BennettNewman.pdf

Newman, D., (2020). 'The Case for a Semantic Data Catalog'. CS 520 Knowledge Graphs - How should AI explicitly represent knowledge? Department of Computer Science, Stanford University, Spring 2020. Retrieved 19 September 2020, from https://web.stanford.edu/class/cs520/abstracts/newman.html

Object Management Group (2019). Semantics of Business Vocabulary and Business Rules, version 1.5. OMG Document Number: formal/2019-10-02. Retrieved on 18 September 2020, from https://www.omg.org/spec/SBVR/1.5/PDF

Odell, J. (2011). Ontology. CSC Catalyst White Paper. Computer Sciences Corporation. Retrieved September 19 2020, from http://www.jamesodell.com/Ontology_White_Paper_2011-07-15.pdf

Ontology Summit (2020). Ontology Summit 2020: Knowledge Graphs. Retrieved 18 September 2020, from https://ontologforum.org/index.php/OntologySummit2020

U.S. Electronic Code of Federal Regulations (2002). Title 12, Chapter II, Subchapter A, Part 223.

Wittgenstein, L. (1953). *Philosophical Investigations.* Macmillan.

World Wide Web Consortium (11 December 2012). OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition). W3C Recommendation. Retrieved 19 September 2020, from https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/

World Wide Web Consortium (25 February 2014). Resource Definition Framework (RDF) Schema 1.1. W3C Recommendation. Retrieved 19 September 2020, from https://www.w3.org/TR/2014/REC-rdf-schema-20140225/

World Wide Web Consortium (20 July 2017). Shapes Constraint Language (SHACL). W3C Recommendation. Retrieved 19 September 2020, from https://www.w3.org/TR/shacl/

Yang, G., Kifer, M., Zhao, C. (2003). FLORA-2: A Rule-Based Knowledge Representation and Inference Infrastructure for the Semantic Web. Lecture Notes in Computer Science 2888:671-688. Springer. DOI: 10.1007/978-3-540-39964-3_43

Yu, J. (2020). Knowledge Graph Use Cases @ Intuit. CS 520 Knowledge Graphs - How should AI explicitly represent knowledge? Department of Computer Science, Stanford University, Spring 2020. Retrieved 19 September 2020, from https://web.stanford.edu/class/cs520/abstracts/yu.pdf

BIO

Mike Bennett is the director of Hypercube Limited, a company that helps people manage their information assets using formal semantics. Mike is the originator of the Financial Industry Business Ontology (FIBO) from the EDM Council, a formal ontology for financial industry concepts and definitions. Mike provides mentoring and training in the application of formal semantics to business problems and strategy, and is retained as Standards Liaison for the EDM Council and the IOTA Foundation, a novel Blockchain-like ecosystem.

Decision Rationales as Models for Explanations

Kenneth Baclawski
Northeastern University

Abstract

A decision rationale describes the reasons for a decision in an engineering or software development process, so it is a kind of explanation. Conversely, explanations are commonly used for decisions that have been made. In this article we develop a reference ontology for decision rationales, which captures the common features of explanations for decisions in a domain-independent manner. The intention is to tie together the many techniques for explainability in different domains so that the techniques can be shared and possibly even interoperate with one another.

Introduction

A DECISION RATIONALE IS AN ARTIFACT that describes the reasons for a decision. In practice organizations commonly do not record the knowledge generated during a decision making process. As a result, it can be an expensive and painful process to revisit a past decision when it becomes apparent that the decision is no longer appropriate (Spacey 2016). This problem has been recognized for software development processes, and there are now a number of software tools that assist developers in capturing and managing decision rationales (See: the Section Decision Rationale Reference Ontology)

It should be apparent that decision rationales are a form of explanation; namely, the answer to why a decision was made. Explanations have recently become an important issue. As stated in the Communiqué of the Ontology Summit (2019), with the increasing amount of software devoted to industrial automation and process control, it is becoming more important than ever for systems to be able to explain their behavior. In some domains, such as financial services, explainability is mandated by law. In spite of this, explanation today is largely handled in an unsystematic manner, if it is handled at all."

While not all explanations are in response to a decision, such explanations are a significant share of all explanations. Accordingly, a framework for decision rationales would contribute to a common framework for explanations in general. In this article we develop a reference ontology for decision rationales. A reference ontology is an intermediate ontology that is more specific than foundational ontologies (also known as "upper ontologies") but more general than domain ontologies. A reference ontology deals with a specific issue but is otherwise domain-independent. The advantage of a reference ontology is that it can link together techniques from different domains for purposes such as data integration, software reuse and interoperability. In particular research and tools for decision rationales for engineering processes could be used for making other systems more explainable.

The reference ontology that we develop originated from the work of Sriram (2002) as well as (Duggar and Baclawski, 2007). The requirements for this ontology were taken from wide range of sources, especially from the Ontology Summit 2019 Baclawski *et al* (2019), and the fields of Explainable Artificial Intelligence Srihari (2020), commonsense knowledge and reasoning Berg-Cross (2020), medical explanations Baker, Al Manir, Brenas, Zinszer, and Shaban-Nejad (2020), and financial explanations (Bennett 2020).

To illustrate a decision making process, we will use a running example of a specific decision making process for dealing with a problem in industry known as No-Trouble-Found (NTF) or No-Faults-Found (Accenture Communications 2016). The NTF problem is that components used in application areas, such as automobiles, electric utilities, and manufacturing, have mechanisms for indicating component failure. The failure is typically advertised with an alarm. When an alarm is raised, the component may be replaced at little or no cost under the terms of a warranty or service contract. The component that raised the alarm is returned to the supplier and tested in their laboratory. Remarkably, as much as 25% to 70% of the time, the returned component operates correctly when tested. To deal with the problem, the manufacturer will need to test the returned components to determine whether they function correctly. This will generally involve a series of tests that are used to make the decision about whether a component is actually faulty as shown in Figure 1. The figure shows a three-step decision making loop, but an actual decision making process for NTF could have many more steps. While the running example we are using is relatively specific, it is similar to many other decision making processes in which there are three alternatives: accept, reject or get more information.

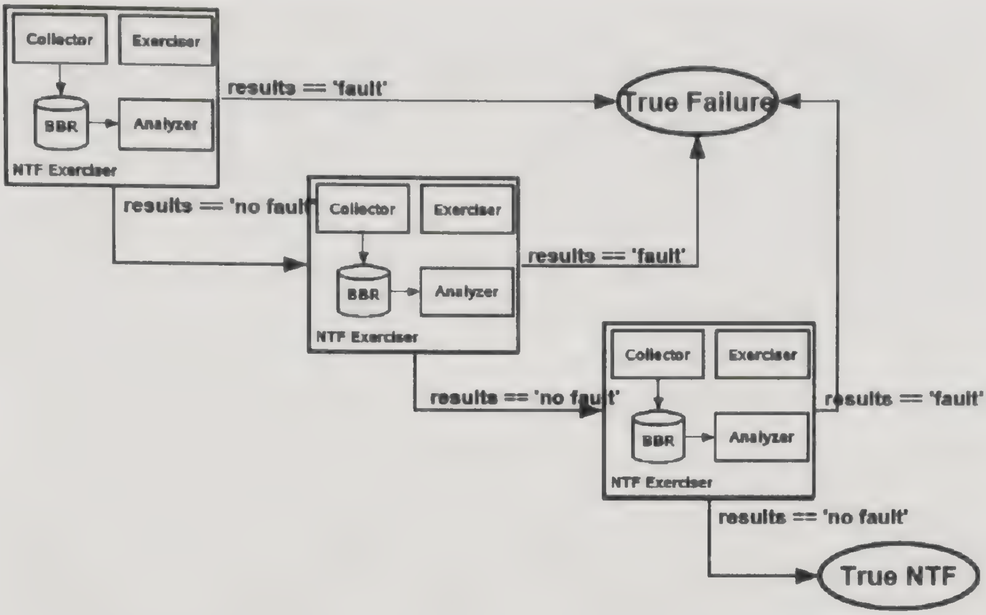


Figure 1: Sequential Hypothesis Flowchart for Electronic Systems and Components Evaluated as "Suspect NTFs" from Baclawski *et al* (2018)

In the Background Section we give some background for decision rationales and compare them with explanations. We then give some of the reasons why it is useful to document decision rationales in the Purpose of Documenting Decision Rationales Section. Generally one must capture decision rationales immediately or not at all. Consequently, decision rationale management must be an integral part of the software development process. Similarly, explainability should drive the software engineering process from the earliest stages of planning, analysis and design (Clancey 2019). In the Decision Rationale Development Process Section we discuss the process whereby decision rationales are developed. The reference ontology decision rationales is presented in the Decision Rationale Reference Ontology Section. We end with a conclusion and acknowledgments.

Background

In this section we give some background on decision rationales and compare them with the more general concept of an explanation. The

Ontology Summit 2019 covered the notion of explanation so it is worthwhile to review the definition of this concept given there:

> An explanation is the answer to the question "Why?" possibly also including answers to follow-up questions such as "Where do I go from here?" Accordingly, explanations generally occur within the context of a process, which could be a dialog between a person and a system or could be an agent-to-agent communication process between two systems. Explanations also occur in social interactions when clarifying a point, expounding a view, or interpreting behavior. In all such circumstances in common parlance one is giving/offering an explanation (Ontology Summit 2019).

While explanation has a long philosophical history dating back at least to 5000 BCE, formal treatments of rationales are relatively recent. Perhaps the earliest such treatment was the school of philosophy known as scholasticism that dominated teaching in European universities from roughly 1100 to 1700. It focused on how to acquire knowledge and how to communicate effectively so that it may be acquired by others. It was thought that the best way to achieve this was by replicating the discovery process and by arguing for and against alternatives (O'Boyle 1998). While scholasticism arose in the context of religious instruction, it soon spread to other disciplines.

Another example of scholarly work on decision rationales is in the legal domain. From ancient times the rationales for legal decisions have been recorded and used in subsequent decisions. This is the basis for what is now referred to as "common law." There is a substantial scholarly literature on decision rationales in the legal domain. This should not be too surprising since argumentation is so fundamental to legal decisions, and since it is still a requirement that not only the decision itself but also the rationale for the decision should be documented.

In spite of rationales being common in the legal domain, they are relatively uncommon in law codes, and even when laws have explicitly stated rationales, their standing is ambiguous. The Constitution of the United States has a published rationale in the form of a series of articles called the Federalist Papers (Hamilton, Madison, and Jay, 1787). However, the Constitution itself does not explicitly include a rationale, so whether the

Federalist Papers could be used by courts for deciding cases is controversial. There is only one amendment, namely the Second Amendment, that explicitly includes a rationale, albeit a very brief one. The interpretation of this rationale and of the amendment as a whole has been highly controversial. Prior to the year 2008 the rationale was taken to be a limitation on the amendment, essentially giving the states the power to organize militias and allowing individuals to bear arms for this purpose. Up to that time states and the federal government had the authority to regulate ownership of arms for other purposes. However, in 2008, the United States Supreme Court reinterpreted the Second Amendment by ignoring the rationale (Brennan Center 2018; Greenhouse 1998). From the second citation: "Many are startled to learn that the U.S. Supreme Court didn't rule that the Second Amendment guarantees an individual's right to own a gun until 2008, when District of Columbia v. Heller struck down the capital's law effectively banning handguns in the home. In fact every other time the court had ruled previously, it had ruled otherwise." This is good case study to show that including or not including a rationale can result in dramatically different interpretations of a law.

Purpose of Documenting Decision Rationales

We now give some of the reasons why decision rationales should be documented and reviewed. Put more succinctly (if somewhat inaccurately), we give a rationale for rationales.

An important part of every decision rationale for software development is the list of the alternatives that were considered. Documenting these options can be useful in themselves. According to Sullivan (1999), "...part of the value of typical software product, process or project is in the form of embedded options. These real options provide design decision-makers with valuable flexibility to change products and plans as uncertainties are resolved over time."

Possibly the most dramatic example of this was a decision for the Ariane V rocket software that was not reconsidered for the Ariane V rocket. The result was that the rocket crashed on its first launch (Gleick 1996).

It might be worth examining in some more detail what the design decision was that resulted in the Ariane V crash. The Ariane V rocket reused vehicle guidance software from the Ariane IV. These different rockets used different processors and the reuse of the Ariane IV software code failed to operate as expected in the Ariane V. The failure occurred in the inertial reference system, or the Système de Référence Inertielle (SRI). The failure was due to a software exception during execution of a data conversion from a 64-bit floating point in a variable for Horizontal Bias (BH) to a 16-bit signed integer value. The floating point number which was converted had a value greater than what could be represented by a 16-bit signed integer. The use of a 16-bit signed integer in the Ariane IV was a design decision that was made during the development of the SRI. This design decision was documented and even rigorously proven to be correct for the Ariane IV. Unfortunately, the specifications for the Ariane IV that were used in this proof are not satisfied by the Ariane V. From the Inquiry Report, "The reason for the three remaining variables, including the one denoting horizontal bias, being unprotected was that further reasoning indicated that they were either physically limited or that there was a large margin of safety, a reasoning which in the case of the variable BH turned out to be faulty. It is important to note that the decision to protect certain variables but not others was taken jointly by project partners at several contractual levels." However, the reasoning (i.e., the proof) was not included in the source code so it was not reviewed when the software was reused for the Ariane V. This is an example to show that a formal proof of correctness of software is useless if it is not reconsidered when circumstances change. It also shows the risks associated with software reuse (Lions 1996).

If, as it is hoped, systems begin to be more explainable, the experiences with rationale management could be useful lessons. As the Ariane V disaster illustrates, one such lesson is the issue of decision rationale reusability. The purpose of reusability is to save time and resources and reduce redundancy by taking advantage of assets that have already been created in some form within the software product development process (Lombard Hill Group 2014). Unfortunately, software reuse has not been very successful in general (Schmidt 1999).

The Decision Rationale Development Process

The process model for decision rationale development is a basic decision making loop, but it extends it by specifying a data model for the resulting rationale. A use case diagram showing two of the actors and activities during formal documentation and use of decision analysis is shown in Figure 2, taken from (Duggar and Baclawski, 2007). The two roles/actors in this figure are the developer of the decision analysis documentation and the user of the decision analysis. The user is the agent who is seeking an explanation of the decision. The developer can perform a number of actions on the repository of decision rationales, such as create, modify, and reuse/repurpose. Other use cases that are not shown are concerned with activities such as reconsidering decisions and inference/reasoning.

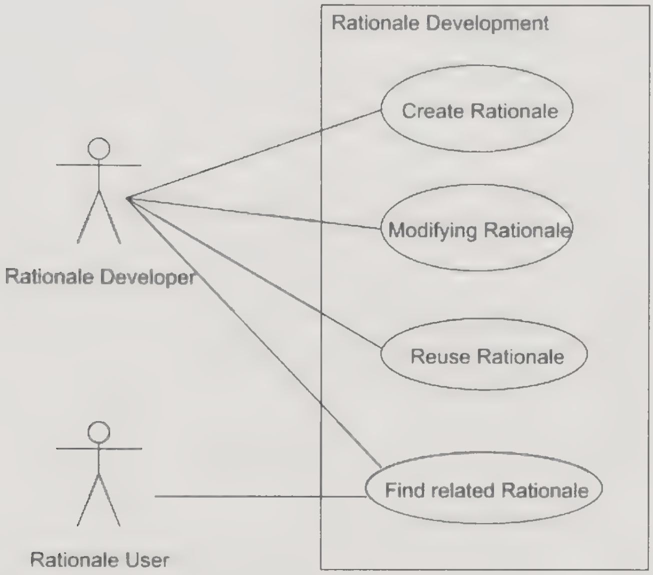


Figure 2: Use Case Diagram for Decision

The process model for decision rationale development is usually a sub-process of a larger development process. When an issue has been encountered for which a decision is required, a decision making process is performed. Determining and identifying the issue to be resolved may itself be a decision that requires its own decision making process. An example of a process for developing the decision rationale is shown in Figure 3. The process involves a number of steps and iterations as follows:

1. Enumerate all the assumptions that are relevant and can be inferred based on the context or situation.
2. Exhaustive list of all the alternatives that can be chosen for a particular decision have to be documented.
3. Similarly, a list of criteria based on which any alternative would be chosen for an issue/problem is documented.

4. Both step 1 and step 2 are done iteratively till a satisfied list of both alternatives and criteria are available.
5. Relevant arguments for each alternative based on the list of criteria are obtained.
6. Based on the arguments put forward a decision is recommended.

The whole process from steps 1 to 6 could then be iterated till a satisfactory decision is obtained.

This decision rationale development loop is a special case of the general decision making loop developed in (Baclawski *et al*, 2017). The ontology for the decision making loop is available online at (Baclawski 2016).

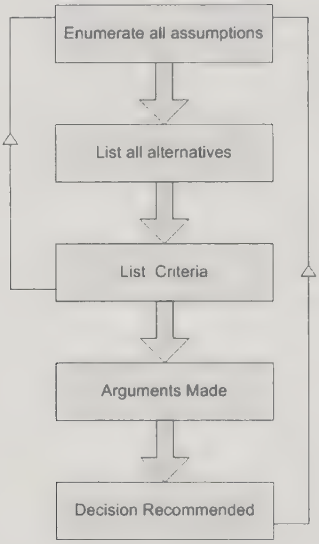


Figure 3: Example of a Decision Rationale Development Loop

For the running example of the NTF decision making process, each step in the process can have three possible outcomes. The component may be found to be actually faulty, the component may be found to be functioning normally, or the component test was unable to establish the condition of the component with sufficient confidence. When the last of these occurs, another test is performed. The rationale for each step in this process has three alternatives. The criterion for each alternative is commonly a range of values for a measurement. The argument for the decision of one step could be as simple as checking whether the measurement is within the range for the corresponding alternative or it could be a more complex statistical or machine learning classification involving both the current measurement and previous measurements.

Some software tools are available for rationale development. Compendium (2020), designVUE (2020) and SEURAT Website (2020) are examples of open source projects that include support for capturing decision rationales. Rationale® (2020) is a commercial product. Gelder (2007) reviews this product. The primary purpose of these tools is to document design decisions during software development. The tools can also be used for documenting more general argumentation, such as in legal cases. These tools do not appear to make use of ontologies.

Decision Rationale Reference Ontology

We now formalize the notion of a decision rationale as a reference ontology. The basis for our ontology is the Design Recommendation and Intent Model (DRIM) that was developed for engineering design decisions but is not limited to that domain (Sriram 2002). The DRIM is shown in Figure 4, using the Object Modeling Technique. We also used some ideas from our own decision rationale ontology in Duggar and Baclawski (2007) which was intended for software engineering using the Eclipse Process Framework.

A number of other reference ontologies were important inputs to our ontology, including reference ontologies for situation awareness, provenance and decision making. Situation awareness means simply that one knows what is going on around oneself. In operational terms, this means that one knows the information that is relevant to a task or goal. The notion of decision rationale fits well with situation awareness, since a decision rationale is the awareness of the information relevant to the making of a decision. Accordingly, we view a decision rationale as a situation. The ontology for situations and situation awareness was first developed in (Baclawski, Malczewski, Kokar, Letkowski, and Matheus, 2002).

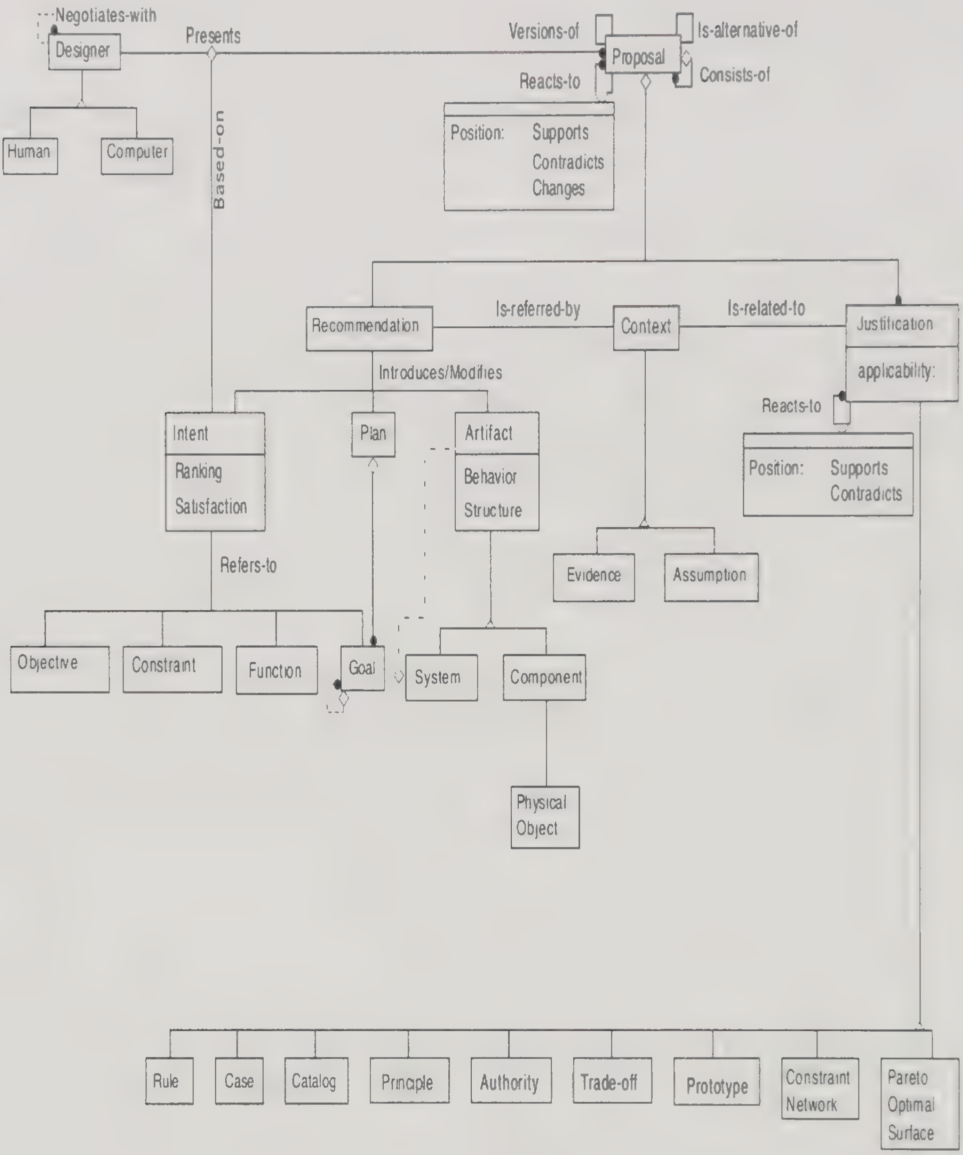


Figure 4: The Design Recommendation and Intent Model

The provenance of an entity represents its origin. This includes descriptions of the other entities and the activities involved in producing and influencing a given entity. All of the various objects involved in a decision rationale are entities for which provenance is important. For example, the decision rationale itself, the problem that is to be solved, the various proposals for solving a problem, and the various arguments in favor of or opposed to each proposal, should all be annotated with the person (or other agent) that created the entity, the time when the entity was created, and so on. The PROV ontology was used for provenance information (PROV Ontology 2013).

Given that a decision rationale is the recording of a decision making process, the process whereby the decision is made should be compatible with the structure of the decision rationale. The ontology for decision making that we use is the Knowledge Intensive Data System (KIDS) (Baclawski *et al*, 2017). In the KIDS framework, the decision making process is a loop in which a situation evolves iteratively to achieve the final decision. In the process, subsidiary decisions will be made, each represented by its own situation. The decision rationale ontology is intended to be one kind of situation that the KIDS framework applies to.

While the decision rationale ontology we present here is intended primarily for software development decision making, it is domain-independent and so has other potential application domains. It could be applied to more general engineering decision making; indeed, this was the original domain for the DRIM model from which the decision rationale ontology was derived. Another potential domain is legal decisions. While we are not aware of any ontologies specifically for legal decision rationales, the legal literature does have examples of work on representing both classifications and argumentation rules (Berman and Hafner, 1993; Loui and Norman, 1995).

The decision rationale ontology was developed using Protégé (Protégé 2004; Musen 2015). It imports the PROV ontology so that provenance information can be maintained in a standard manner PROV (2013), and all of the decision rationale ontology classes are subclasses of the prov:Entity class, except for the Collaboration class which is a subclass of prov:Activity. The Rationale class is a subclass of the kids:DirectiveSituation class of the KIDS ontology Baclawski *et al* (2017) which, in turn, is a subclass of the sto:Situation class of the Situation Theory Ontology (Baclawski, Malczewski, Kokar, Letkowski, and Matheus, 2006). The Decision Rationale Ontology is available online (Baclawski 2020).

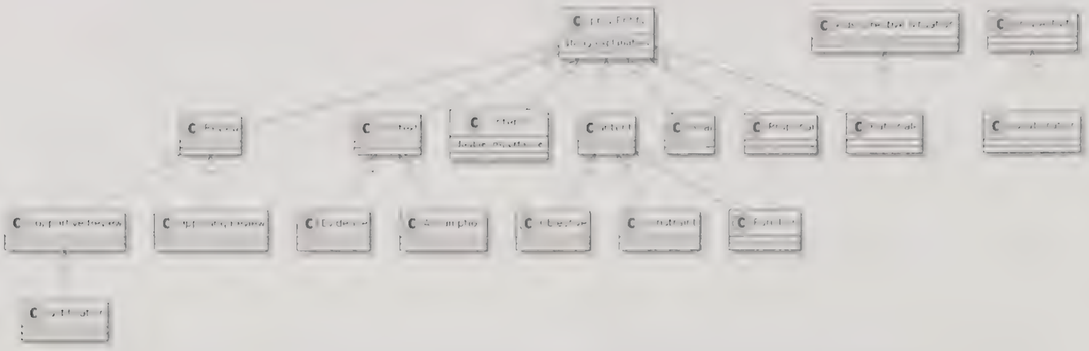


Figure 5: Class Hierarchy of the Decision Rationale Ontology

Figure 5 shows the class hierarchy of the Decision Rationale Ontology. The notation in this figure uses a UML-like notation, but the classes in the hierarchy are not limited to classes in object-oriented software engineering. As noted earlier, all of the classes are subclasses of either prov:Entity or prov:Activity. As a result, all decision rationale artifacts will have all of the many features that the PROV ontology provides, including versioning and provenance information. We added an explanation attribute to the prov:Entity class so that all decision rationale artifacts have a uniform way to explain their role. Potentially, the explanation attribute could contribute to an explainability process as discussed below. The explanation attribute is specified to be a string, but other media could also be used such as diagrams or videos.

The central class in the ontology is the Rationale class. This class reifies the notion of a decision rationale. In addition to being a subclass of prov:Entity, the Rationale class is a subclass of kids:DirectiveSituation which links the Rationale with the ontology of the decision making process that produces the decision rationale. During such a process, a decision may depend on other decisions, and this is represented by the dependsOn object property. In the running example of an NTF decision making process, each step of the process depends on the previous step. To understand how the Rationale class represents a decision rationale we need to examine the object properties shown in Figure 6. The Goal class represents the problem that the decision process is solving. For the NTF problem the goal is to determine whether or not a returned component is actually faulty. Various alternative Proposals are suggested, one of which is selected as the recommendation. A Proposal can have sub-proposals specified by the consistsOf object property.

In the NTF example, the three alternatives are the proposals. In this example, a proposal could have a more complex structure if a component test is more elaborate. Indeed, a component test could itself be a decision making process. The proposals need to satisfy various Criteria. The Criterion class serves to specify how important a particular requirement is, where an importance level of 1 means that the criterion is mandatory, while lower levels represent criteria that are desirable but not essential. The actual requirement is specified by the Intent class, which has subclasses Objective, Constraint, and Function that specify different kinds of requirements. An Objective is a characteristic that is to be optimized. A Constraint is a mandatory restriction such as a maximum allowed value. A Function requirement is a performance characteristic of activities or behavior of the solution to the problem. For the NTF example, the Intent could be a Constraint if the test is a simple measurement or the Intent could be a Function if the test is a more complex test of the behavior of the component.

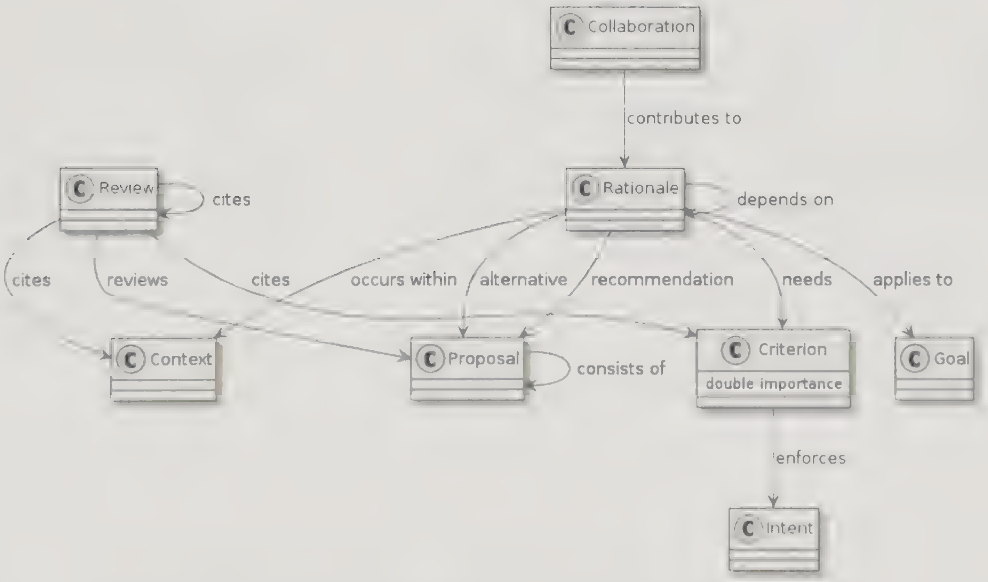


Figure 6: Object Properties of the Decision Rationale Ontology

Since the decision is a selection among alternatives, one needs some way to distinguish them. This is done by means of Reviews. Each Review gives an argument either in favor of a Proposal or against a Proposal. The subclasses SupportiveReview and OpposingReview distinguish these two cases. The explanation for the final decision that selects the recommendation is the Justification. Since a Justification is supportive of the recommended Proposal, Justification is a subclass SupportiveReview. Reviews can cite other Reviews and Criteria in their explanation. A Review can also cite Context information. Context represents background information that may be relevant to the decision. There are two subclasses of Context. The Evidence subclass represents observations and experiments that are relevant to the decision process and believed to be facts. The Assumption subclass represents conjectural information that may or may not be the case but which is relevant to the decision process. Context information may include references to published research papers or books.

Reviews can be the result of a collaborative process involving several individuals. Such a process could be cooperative or antagonistic. If the latter, then the collaboration is likely to represent a negotiation process. At first it appears that Collaboration is unconnected with any Reviews or individuals. In fact, there are connections, but they are represented using object properties of the PROV ontology, and so do not appear in Figure 6.

The Decision Rationale Ontology is derived from the DRIM model shown in Figure 4. Most of the classes in the Decision Rationale Ontology have the same (or very close) meaning as the corresponding class in DRIM. The Context, Evidence, Assumption, Objective, Constraint, Function, Goal and Proposal are the same as in DRIM. For more about what these classes mean, see Chapter 8 of (Sriram 2002). The Review class hierarchy was derived from the Justification and Recommendation classes in DRIM. For example, Recommendation has been replaced by the recommendation object property, but the meaning is largely the same. The Decision Rationale Ontology reifies as classes some characteristics of DRIM that were not classes or were implicit. The negotiates-with relationship is reified as the Collaboration class, which allows the collaboration to be a subclass of prov:Activity. The relationships with Intent were reified so that they could have additional information. The Intent class itself differs only in that Goal is no longer a subclass. This was done to make it easier to integrate the ontology with decision making ontologies such as KIDS. The Designer class of DRIM is represented with the prov:Agent class. The versions-of and is-alternative-to object properties are represented with the prov:wasDerivedFrom, prov:alternateOf, and prov:specializationOf object properties of PROV, although the meanings are somewhat different. The Plan, Artifact and Physical Object classes of DRIM were not included because they deal with the subsequent implementation of the decision, which is important but out of scope to the decision rationale.

There are several steps in formulating any explanation about a system. As noted above, an explanation is the answer to the question "Why?" possibly also including answers to follow-up questions. The goal that is being achieved by the decision rationale should be explained sufficiently so that one can find the decision rationales that are relevant to the question by using search techniques. The explanations associated with each entity in a decision rationale may be used to answer questions about the decision rationale. The justification of a decision rationale explains why the decision proposal was selected (*i.e.*, the answer to a question about why the recommended proposal was chosen). For the running example of the NTF decision making process, the explanation for why the component was either put back in the warehouse or thrown away is in the Justification of the final recommendation of the process. Other reviews explain why alternative proposals were not selected (*i.e.*, the answer to a counter-factual question about why another proposal was not chosen). In the NTF example, one might ask why further testing was not performed. This would be especially important if the component was very expensive. The criteria that constrain the potential proposals explain why other possibilities were not considered (*i.e.*, the answer to contrastive questions about why another decision was not considered). In the NTF example, one might ask why the customer who returned the component was not contacted to determine more information about why the component was thought to be faulty. The explanation is simply that the goal was only to determine whether the component was faulty, not why it was returned. The dependencies among decision rationales allow for follow-up questions that explore decisions in more depth. In the NTF example, one might inquire about the reason for the goal or why the tests were being performed in the particular order and not some other order. These are concerned with the design of the process rather than the process steps. The design was the result of its own decision making process and rationale. An example of how one can optimize the order of the steps in the NTF decision making process is developed in Section II of (Baclawski *et al*, 2018).

Conclusion

We have shown that decision rationales are an effective basis for explaining some features of a system. Specifically, we have shown how decision rationales can be used to answer all of the main kinds of explanation questions for decisions: direct questions, counter-factual questions, contrastive questions, and followup questions. We also discussed how decision rationales can be developed, and presented a reference ontology for decision rationales. Having explored the concept of the decision rationale, we propose that they could be a significant contributor to explainability.

Acknowledgments

This work was conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

References

Baclawski, K. (2016). The KIDS Ontology version 2.0. Retrieved from http://bit.ly/2xZuTNJ

Baclawski, K. (Ed.). (2019). Ontology Summit 2019: Explanations. Retrieved from http://bit.ly/2z0JGY4

Baclawski, K. (2020). Rationale ontology. Retrieved from https://bit.ly/3eg4HQO

Baclawski, K., Bennett, M., Berg-Cross, G., Fritzsche, D., Sharma, R., Singer, J., . . . Whitten, D. (2020). Ontology Summit 2019 Communiqué: Explanation. Applied Ontology. DOI: 10.3233/AO-200226

Baclawski, K., Chan, E., Gawlick, D., Ghoneimy, A., Gross, K., Liu, Z., & Zhang, X. (2017). Framework for ontology-driven decision making. Applied Ontology, 12 (3-4), 245–273. Retrieved from https://bit.ly/2LYPszt

Baclawski, K., Chystiakova, A., Gross, K., Gawlick, D., Ghoneimy, A., & Liu, Z. (2019, April). Use cases for machine-based situation awareness evaluation. In IEEE Conference on Cognitive and Computational Aspects of Situation Management.

Baclawski, K., Malczewski, M., Kokar, M., Letkowski, J., & Matheus, C. (2002, November 4). Formalization of situation awareness. In Eleventh OOPSLA Workshop on Behavioral Semantics (pp. 1–15). Seattle, WA.

Baclawski, K., Malczewski, M., Kokar, M., Letkowski, J., & Matheus, C. (2006). The Situation Theory Ontology. Retrieved from http://bit.ly/1yrikQj

Baker, C., Al-Manir, M., Brenas, J., Zinszer, K., & Shaban-Nejad, A. (2020). Applied Ontologies for Global Health Surveillance and Pandemic Intelligence. J. Wash. Acad. Sci. **106**, No. 4

Bennett, M. (2020). Financial Industry Explanations. J. Wash. Acad. Sci. **106**, No. 4

Berg-Cross, G. (2020). Commonsense and Explanation: Synergy and Challenges in the Era of Deep Learning Systems. J. Wash. Acad. Sci. **106**, No. 4

Berman, D., & Hafner, C. (1993). Representing teleological structure in case-based legal reasoning: The missing link. In Fourth Int. Conf. On Artificial Intelligence and Law (pp. 50–59).

Big Trouble with "No Trouble Found" Returns: Confronting the High Cost of Customer Returns. (2016). Retrieved from http://bit.ly/2vO5QZD

Clancey, W. (2019, February). Explainable AI Past, Present, and Future: A Scientific Modeling Approach. Retrieved from http://bit.ly/2Scjvo6

The Compendium Website. (n.d.). Retrieved from http://bit.ly/3avsavd

The designVUE Website. (n.d.). Retrieved from http://bit.ly/2yFGpA3

Duggar, V., & Baclawski, K. (2007, November 5). Integration of decision analysis in process life-cycle models. In International Workshop on Living with Uncertainties. Atlanta, Georgia, USA.

Gleick, J. (1996, December 1). A bug and a crash: Sometimes a bug is more than a nuisance. New York Times Magazine.

Greenhouse, L. (27, 2008, June). Justices, ruling 5-4, endorse personal right to own gun. In The New York Times. Retrieved from https://nyti.ms/3giVER3

Hamilton, A., Madison, J., & Jay, J. (1787). The Federalist: a Collection of Essays, Written in Favour of the New Constitution, as Agreed upon by the Federal Convention, September 17, 1787, in two volumes (1st ed.). New York: J. and A. McLean.

How the NRA Rewrote the Second Amendment - Brennan Center for Justice. (2018). Retrieved from https://bit.ly/2zwypCq

Lebo, T., Sahoo, S., & McGuinness, D. (Eds.). (2013, April 30).

PROV Ontology (PROV-O). Retrieved from http://bit.ly/2xPcx2k

Lions, J. (1996). ARIANE 5 Flight 501 Failure: Report by the Inquiry Board. Retrieved from https://bit.ly/2ZzUoDb

Loui, R., & Norman, J. (1995). Rationales and argument moves. Artif. Intell. Law , 3 , 159–189. Retrieved from https://doi.org/10.1007/BF00872529

Musen, M. (2015, June). The Protégé project: A look back and a look forward. In AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence (Vol. 1). Retrieved from https://bit.ly/2zwqBQY

O'Boyle, C. (1998). The art of medicine: medical teaching at the University of Paris, 1250-1400. Leiden: Brill.

Protégé website. (2004). Retrieved from http://bit.ly/AASA

The Rationale ® Website. (n.d.). Retrieved from https://bit.ly/2TKLgla

Schmidt, D. (1999). Why software reuse has failed and how to make it work for you (Tech. Rep.). Nashville, Tennessee: Vanderbilt University. Retrieved from https://bit.ly/2M1vGUc

The SEURAT Website. (n.d.). Retrieved from http://bit.ly/2VPfLg3

Spacey, J. (2016). What is a design rationale? Retrieved from https://bit.ly/3c5vf5V

Srihari, S. (2020). Explainable Artificial Intelligence: An Overview. J. Wash. Acad. Sci. **106**, No. 4, p. 1

Sriram, R. (2002). Distributed and Integrated Collaborative Engineering Design. Glenwood, MD 21738: Sarven Publishers. ISBN 0-9725064-0-3

Sullivan, K. (1999). Software design as an investment activity: A real options perspective. Real Options and Business Strategy: Applications to Decision Making, 215–261.

van Gelder, T. (2007). The rationale for Rationale ®. Law, Probability and Risk, 6, 23–42. Retrieved from doi:10.1093/lpr/mgm032

What Is Software Reuse? (2014). Lombard Hill Group. Retrieved from http://www.lombardhill.com

BIO

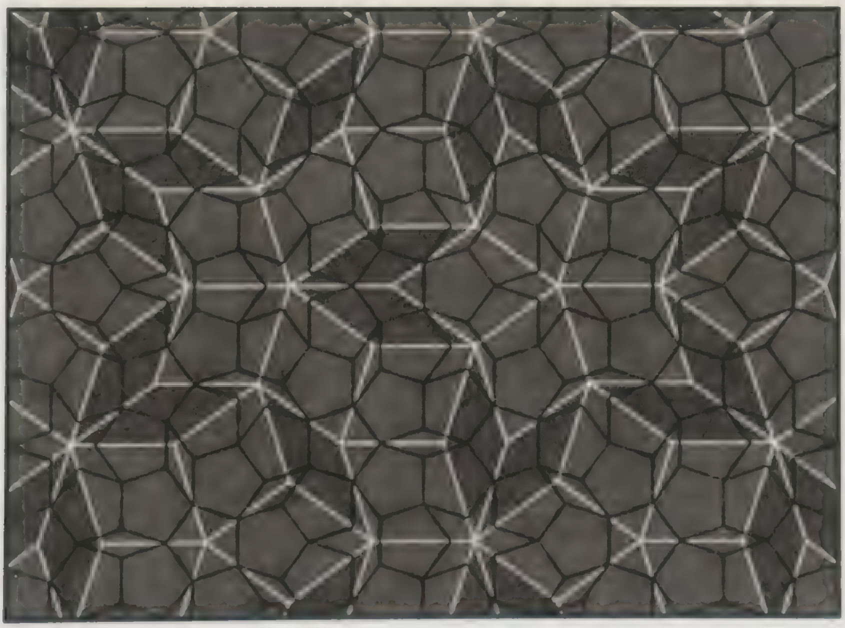
Kenneth Baclawski is an Associate Professor Emeritus at the College of Computer and Information Science, Northeastern University. Professor Baclawski does research in data semantics, formal methods for software engineering and software modeling, data mining in biology and medicine, semantic collaboration tools, situation awareness, information fusion, self-aware and self-adaptive systems, and wireless communication. He is a member of the Washington Academy of Sciences, IEEE, ACM, IAOA, and is the chair of the Board of Trustees of the Ontolog Forum.

Science Bite: Sir Roger Penrose and Penrose Tilings

Sir Roger Penrose used clever mathematical arguments in 1965 to prove that black holes are a direct consequence of Albert Einstein's general theory of relativity. This is considered the most important contribution to the general theory of relativity since Einstein. For this result he received many awards including most recently a 50% share of the 2020 Nobel Prize in Physics. He is Emeritus Rouse Ball Professor of Mathematics at the University of Oxford.

For this science bite we highlight a mathematical achievement by Penrose not related to astrophysics. This is the construction in 1974 of Penrose Tilings. Tilings are collections of pieces, typically polygons, that fill up the plane with no gaps and no overlaps. It was thought that tilings had to be periodic, that is, have a pattern that repeats itself over and over. Penrose constructed aperiodic tilings, which are formed from two tiles that can only tile the plane non-periodically. Although a result in pure mathematics, these tilings turned out to be closely related to quasicrystals in chemistry. In 1984 such patterns were observed in the arrangement of atoms in quasicrystals (ordered but not periodic).

Below we show one example of a Penrose tiling:



(By Inductive load, in public domain)

Membership List

ABEL, DAVID (Dr.) 14005 Youderian Drive, Bowie MD 20721 (LM)

AKSYUK, VLADIMIR A. 605 Gatestone Mews, Gaithersburg MD 20878 (F)

ANTMAN, STUART (Dr.) University of Maryland, 2309 Mathematics Building, College Park MD 20742-4015 (EF)

APPETITI, EMANUELA Botany Center, The Huntington, 1151 Oxford Road, San Marino CA 91108 (LM)

APPLE, DAINA DRAVNIEKS (Mrs.) PO Box 905, Benicia CA 94510-0905 (M)

ARSEM, COLLINS (Mr.) 3144 Gracefield Rd Apt 117, Silver spring MD 20904-5878 (EM)

ARVESON, PAUL T. (Mr.) 6902 Breezewood Terrace, Rockville MD 20852-4324 (F)

BACLAWSKI, KENNETH (Dr.) 35 Fairmont Ave., Waltham MA 02453 (M)

BARBOUR, LARRY L. (Mr.) Pequest Valley Farm, 585 Townsbury Road, Great Meadows NJ 07838 (M)

BARWICK, W. ALLEN (Dr.) 13620 Maidstone Lane, Potomac MD 20854-1008 (F)

BASCH, PETER (Dr) 5824 Chevy Chase Parkway, NW, Washington DC 20015 (F)

BECKER, EDWIN D. (Dr.) 339 Springvale Road, Great Falls Va 22066 (EF)

BEHLING, NORIKO (Ms.) 6517 Deidre Terrace, McLean VA 22101 (M)

BERRY, JESSE F. (Mr.) 2601 Oakenshield Drive, Rockville MD 20854 (M)

BIONDO, SAMUEL J. (Dr.) P.O. Box 226, Damascus MD 20872 (EF)

BOISVERT, RONALD F. (Dr.) Mail Stop 8910, NIST, 100 Bureau Drive, Gaithersburg MD 20899-8910 (F)

BOSSE, ANGELIQUE P 11700 Stonewood Lane, Rockville MD 20852 (F)

BRADY, KATHIE (Ms.) 4539 Metropolitan Court, Frederick MD 21704 (M)

BRISKMAN, ROBERT D. (Mr.) 61 Valerian Court, North Bethesda MD 20852 (EF)

BROWN, ELISE A.B. (Dr.) 6811 Nesbitt Place, Mclean VA 22101-2133 (LF)

BUFORD, MARILYN (Dr.) P.O. Box 171, Pattison TX 77466 (EF)

BULLARD, JEFFREY WAYNE (Dr.) 2600 Goodrich Court, College Station TX 77845 (F)

BYRD, GENE GILBERT (Dr.) Box 1326, Tuscaloosa AL 35403 (M)

CARASSO, ALFRED (Mr.) 18 War Admiral Court, North Potomac MD 20878 (F)

CAVINATO, TIZIANA (Dr) FCC, 7932 Opossumtown Pike, Frederick MD 21702 (M)

CERRONI, PHILIP VINCENT (Mr.) 5340 Greene St., 1F, Philadelphia PA 19144 (M)

CIORNEIU, BORIS (Dr.) 20069 Great Falls Forest Dr., Great Falls VA 22066 (M)

COBLE, MICHAEL (Dr.) UNTHSC, 3500 Camp Bowie Blvd, CBH-445, Ft. Worth TX 76107 (F)

COFFEY, TIMOTHY P. (Dr.) 976 Spencer Rd., McLean VA 22102 (F)

COHEN, MICHAEL P. (Dr.) 1615 Q. St. NW T-1, Washington DC 20009-6310 (LF)

COLE, JAMES H. (Mr.) 9709 Katie Leigh Ct, Great Falls VA 22066-3800 (F)

CROSS, SUE (Dr.) 12377 Elm Ridge Lane, Ashland Va 23005 (M)

CUPERO, JERRI ANNE (Dr.) 2860 Graham Road, Falls Church VA 22042 (F)

DANNER, DAVID L. (Dr.) 1364, Suite 101, Beverly Road, McLean VA 22101 (F)

DEAN, DONNA (Dr.) 367 Mound Builder Loop, Hedgesville WV 25427-7211 (EF)

DEDRICK, ROBERT L. (Dr.) 21 Green Pond Rd, Saranac Lake NY 12983 (EF)

DHARKAR, POORVA (Dr.) 263 Congressional Lane Apt 412, Rockville MD 20852 (M)

DIMITOGLOU, GEORGE (Dr.) 11053 Seven Hill Lane, Potomac MD 20854 (M)

DONALDSON, JOHANNA B. (Mrs.) 3020 North Edison Street, Arlington VA 22207 (EF)

DOYLE, ELIZABETH K 6705B Overton Circle Apt. 16, Frederick MD 21703 (M)

DURRANI, SAJJAD (Dr.) 17513 Lafayette Dr, OLNEY MD 20832 (EF)

EDINGER, STANLEY EVAN (Dr.) Apt #1016, 5801 Nicholson Lane, North Bethesda MD 20852 (EM)

EGENREIDER, JAMES A. (Dr.) 1615 North Cleveland Street, Arlington VA 22201 (LF)

ERICKSON, TERRELL A. (Ms.) 4806 Cherokee St., College Park MD 20740-1865 (M)

ETTER, PAUL C. (Mr.) 8612 Wintergreen Court, Unit 304, Odenton MD 21113 (F)

FASANELLI, FLORENCE (Dr.) 4711 Davenport Street, Washington DC 20016 (EF)

FASOLKA, MICHAEL J. NIST Material Measurement Laboratory, MS8300, 100 Bureau Dr., Gaithersburg MD 20809 (F)
FAULKNER, JOSEPH A. (Mr.) 2 Bay Drive, Lewes DE 19958 (EF)
FRASER, GERALD (Dr.) 5811 Cromwell Drive, Bethesda MD 20816 (M)
FREEMAN, ERNEST R. (Mr.) 5357 Strathmore Avenue, Kensington MD 20895-1160 (LEF)
FREHILL, LISA (Dr.) 1239 Vermont Ave NW #204, Washington DC 20005-3643 (M)
FROST, HOLLY C. (Dr.) 5740 Crownleigh Court, Burke VA 22015 (F)
GAGE, DOUGLAS W. (Dr.) XPM Technologies, 1020 N. Quincy Street, Apt 116, Arlington VA 22201-4637 (M)
GARFINKEL, SIMSON L. (Dr.) 1186 N Utah Street, Arlington VA 22201 (M)
GAUNAURD, GUILLERMO C. (Dr.) 4807 Macon Road, Rockville MD 20852-2348 (EF)
GHARAVI, HAMID (Dr.) NIST, MS 8920, Gaithersburg MD 20899-8920 (F)
GIBBON, JOROME (Mr.) 311 Pennsylvania Avenue, Falls Church VA 22046 (F)
GIFFORD, PROSSER (Dr.) 59 Penzance Rd, Woods Hole MA 02543-1043 (EF)
GLUCKMAN, ALBERT G. (Mr.) 18123 Homeland Drive, Olney MD 20832-1792 (EF)
GRAY, JOHN E. (Mr.) PO Box 489, Dahlgren VA 22448-0489 (M)
GRAY, MARY (Professor) Department of Mathematics & Statistics, American University, 4400 Massachusetts Avenue NW, Washington DC 20016-8050 (F)
GRUIS, LESLIE (Dr.) 13724 Canal Vista Ct., Potomac Maryland 20854 (M)
HACK, HARVEY (Dr.) 176, Via Dante, Arnold MD 21012-1315 (F)
HAIG, SJ, FRANK R. (Rev.) Loyola University Maryland, 4501 North Charles St, Baltimore MD 21210-2699 (EF)
HARDIS, JONATHAN E. (Dr.) 356 Chestertown St., Gaithersburg MD 20878-5724 (F)
HAYNES, ELIZABETH D. (Mrs.) 7418 Spring Village Dr., Apt. CS 422, Springfield VA 22150-4931 (EM)
HAZAN, PAUL 14528 Chesterfield Rd, Rockville MD 20853 (F)
HEANEY, JAMES B. 6 Olivewood Ct, Greenbelt MD 20770 (M)
HERBST, ROBERT L. (Mr.) 4109 Wynnwood Drive, Annandale VA 22003 (LF)
HIETALA, RONALD (Dr.) 6351 Waterway Drive, Falls Church VA 22044-1322 (M)
HOFFELD, J. TERRELL (Dr.) 11307 Ashley Drive, Rockville MD 20852-2403 (F)
HORLICK, JEFFREY (Mr.) 8 Duvall Lane, Gaithersburg MD 20877-1838 (F)
HORN, JOANNE (Dr) 1408 Grouse Court, Frederick MD 21701 (M)
HOWARD, SETHANNE (Dr.) Apt 311, 7570 Monarch Mills Way, Columbia MD 21046 (LF)
HOWARD-PEEBLES, PATRICIA (Dr.) 5701 Virginia Parkway 2312, McKinney TX 75071 (EF)
IKOSSI, KIKI (Dr.) 6275 Gentle LN, Alexandria VA 22310 (F)
IZADJOO, MINA (Dr.) 15713 Thistlebridge Drive, Rockville MD 20853 (F)
IZADJOO, PARVIS 15713 Thistlebridge Drive, Rockville MD 20853 (M)
JOHNSON, EDGAR M. (Dr.) 1384 Mission San Carlos Drive, Amelia Island FL 32034 (LF)
JOHNSON, GEORGE P. (Dr.) 3614 34th Street, N.W., Washington DC 20008 (EF)
JOHNSON, JEAN M. (Dr.) 3614 34th Street, N.W., Washington DC 20008 (EF)
JONG, SHUNG-CHANG (Dr.) 8892 Whitechurch Ct, Bristow VA 20136 (LF)
KAHN, ROBERT E. (Dr.) 909 Lynton Place, Mclean VA 22102 (F)
KAPETANAKOS, C.A. (Dr.) 4431 MacArthur Blvd, Washington DC 20007 (EF)
KARAM, LISA (Dr.) 8105 Plum Creek Drive, Gaithersburg MD 20882-4446 (F)
KAUFHOLD, JOHN (Dr.) Suite 1200, 4601 N. Fairfax Dr, Arlington VA 22203 (LF)
KEISER, BERNHARD E. (Dr.) 2046 Carrhill Road, Vienna VA 22181-2917 (LF)
KLINGSBERG, CYRUS (Dr.) Apt. L184, 500 E. Marylyn Ave, State College PA 16801-6225 (EF)
KLOPFENSTEIN, REX C. (Mr.) 4224 Worcester Dr., Fairfax VA 22032-1140 (LF)
KOWTHA, VIJAYANAND (Dr) 8009 Craddock Road, Greenbelt MD 20770 (F)
KRUEGER, GERALD P. (Dr.) Krueger Ergonomics Consultants, 4105 Komes Court, Alexandria VA 22306-1252 (EF)
LABOV, JAY B. Keck Center Room 638, 500 Fifth Street, NW, Washington DC 20001 (F)
LAWSON, ROGER H. (Dr.) 10613 Steamboat Landing, Columbia MD 21044 (EF)
LEIBOWITZ, LAWRENCE M. (Dr.) 2905 Saintsbury Plaza, #217, Fairfax VA 22031-1164 (LF)

LEMKIN, PETER (Dr.) 148 Keeneland Circle, North Potomac MD 20878 (EM)
LESHUK, RICHARD (Mr.) 9004 Paddock Lane, Potomac MD 20854 (M)
LEWIS, DAVID C. (Dr.) 27 Bolling Circle, Palmyra VA 22963 (F)
LIBELO, LOUIS F. (Dr.) 9413 Bulls Run Parkway, Bethesda MD 20817 (LF)
LIDDLE, J ALEXANDER (Dr) NIST, MS 6203, 100 Bureau Drive, Gaithersburg MD 20899-6200 (F)
LONDON, MARILYN (Ms.) 3520 Nimitz Rd, Kensington MD 20895 (EF)
LONGSTRETH, III, WALLACE I (Mr.) 8709 Humming Bird Court, Laurel MD 207231254 (EM)
LOOMIS, TOM H. W. (Mr.) 11502 Allview Dr., Beltsville MD 20705 (EM)
LOZIER, DANIEL W (Dr.) 5230 Sherier Place NW, Washington DC 20016 (F)
LUTZ, ROBERT J. (Dr.) 6031 Willow Glen Dr, Wilmington NC 28412 (EF)
LYONS, JOHN W. (Dr.) 7430 Woodville Road, Mt. Airy MD 21771 (EF)
MANDERSCHEID, RONALD W. (Dr.) 10837 Admirals Way, Potomac MD 20854-1232 (LF)
MANI, MAHESH (Dr.) 210 Summit Hall Rd, Gaithersburg MD 20877 (M)
MANOCHA, DINESH 8125 Paint Branch Drive, #5164, College Park MD 20742 (F)
MARRETT, CORA (Dr.) 7517 Farmington Way, Madison WI 53717 (EF)
MATHER, JOHN (Dr.) 3400 Rosemary Lane, Hyattsville MD 20782 (F)
MCFADDEN, GEOFFREY B (Dr.) 100 Bureau Drive, Stop 8910, Gaithersburg MD 20899 (F)
MCGRATTAN, KEVIN B. (Dr.) 11512 Brandy Hall Lane, Gaithersburg MD 20878 (F)
MCNEELY, CONNIE L. (Dr.) School of Public Policy, George Mason University, 3351 Fairfax Dr Stop 3B1, Arlington VA 22201 (M)
MEISAMI, PARISA 139 Lamont Ln, Gaithersburg MD 20878 (M)
MENZER, ROBERT E. (Dr.) 90 Highpoint Dr, Gulf Breeze FL 32561-4014 (EF)
MESSINA, CARLA G. (Mrs.) 9800 Marquette Drive, Bethesda MD 20817 (EF)
METAILIE, GEORGES C. (DR.) 18 Rue Liancourt, 75014 Paris , FRANCE (F)
MIGLER, KALMAN B. (Dr.) NIST, 100 Bureau Drive, Stop 8542, Gaithersburg, MD 20899 (F)
MILLER, JAY H. (Mr.) 8924 Ridge Place, Bethesda MD 20817-3364 (M)
MILLER II, ROBERT D. (Dr.) The Catholic University of America, 10918 Dresden Drive, Beltsville MD 20705 (M)
MOAYERI, NADER (Dr) 9329 Sprinklewood Lane, Potomac MD 20854-2259 (F)
MOUNTAIN, RAYMOND D. (Dr.) 701 King Farm Blvd #327, Rockville MD 20850 (F)
MUMMA, MICHAEL J. (Dr.) 640 Willow Valley Sq., Apt I-509, Lancaster PA 17602-4870 (F)
MURDOCH, WALLACE P. (Dr.) 65 Magaw Avenue, Carlisle PA 17015 (EF)
NEUBAUER, WERNER G. (Dr.) Apt 349, 7820 Walking Horse Circle, Germantown TN 38138 (EF)
NOE, ADRIANNE (Dr.) 9504 Colesville Road, Silver Spring MD 20901 (F)
O'HARE, JOHN J. (Dr.) 108 Rutland Blvd, West Palm Beach FL 33405-5057 (EF)
OHRINGER, LEE (Mr.) 5014 Rodman Road, Bethesda MD 20816 (EF)
OTT, WILLIAM R (Dr.) 19125 N. Pike CreekPlace, Montgomery Village MD 20886 (EF)
PARR, ALBERT C (Dr.) 2656 SW Eastwood Avenue, Gresham OR 97080-9477 (F)
PAULONIS, JOHN J (Mr.) P.O. Box 703, Mohegan Lake NY 10547 (M)
PAZ, ELVIRA L. (Dr.) 172 Cook Hill Road, Wallingford CT 06492 (LEF)
PERSILY, ANDREW K NIST, Mailstop 8630, 100 Bureau Drive, Gaithersburg MD 20899 (F)
PICKHOLTZ, RAYMOND L. (Dr.) 3613 Glenbrook Road, Fairfax VA 22031-3210 (EF)
PLESNIAK, MICHAEL W. 1400 Laurel Dr., Accokeek MD 20607 (F)
POLAVARAPU, MURTY 10416 Hunter Ridge Dr., Oakton VA 22124 (LF)
POLINSKI, ROMUALD (Dr) Prof, Doctor of Sciences (Economics), Ul. Generala Bora 39/87, 03-982 WARSZAWA 131 , Poland (M)
PRZYTYCKI, JOZEF H. (Prof.) 10005 Broad St, Bethesda MD 20814 (LF)
PYKE, JR, THOMAS N. (Mr.) 4887 N. 35th Road, Arlington VA 22207 (EF)
RAMAIAH, MALA (Dr.) 417 Christopher Avenue Apt. 24, Gaithersburg MD 20879 (M)
RANSOM, BARBARA (Dr.) 3117 8th North, Arlington 22201 (M)

REGLI, WILLIAM (Dr) Department of Computer Science, Institute for Systems Research, Clark School of Engineering, 2173 A.V. Williams Building, 8223 Paint Branch Drive, University of Maryland, College Park MD 20742 (F)

REISCHAUER, ROBERT (Dr.) 5509 Mohican Rd, Bethesda MD 20816 (EF)

RICKER, RICHARD (Dr.) 12809 Talley Ln, Darnestown MD 20878-6108 (EF)

RIDGELL, MARY P.O. Box 133, 48073 Mattapany Road, St. Mary's City MD 20686-0133 (LM)

ROBERTS, SUSAN (Dr.) Ocean Studies Board, Keck 607, National Research Council, 500 Fifth Street, NW, Washington DC 20001 (F)

ROGERS, KENNETH (Dr.) 355 Fellowship Circle, Gaithersburg MD 20877 (LM)

ROMAN, NANCY GRACE (Dr.) 8100 Connecticut Ave. Apt.1605, Chevy Chase MD 20815 (M)

ROMIG, JR, ALTON National Academy of Engineering, 2101 Constitution Ave., NW, Washington DC 20001 (F)

ROOD, SALLY A (Dr.) PO Box 12093, Arlington VA 22219 (F)

ROSENBLATT, JOAN R. (Dr.) 701 King Farm Blvd, Apt 630, Rockville MD 20850 (M)

SANDERS, JAY (Dr.) 7850 Westmont Lane, McLean VA 22102 (F)

SAUBERMAN, P.E., HARRY R (Mr.) 8810 Sandy Ridge Ct., Fairfax VA 22031 (M)

SAUNDERS, BONITA V. 131 Goucher Terrace, Gaithersburg MD 20877 (F)

SCHMEIDLER, NEAL F. (Mr.) 7218 Hadlow Drive, Springfield VA 22152 (F)

SELKIRK, WILLIAM 2423 Wynfield Ct, Frederick MD 21702 (M)

SENKEVITCH, EMILEE (Dr) 1015 Columbine Drive, Apt 2B, Frederick MD 21701 (M)

SEVERINSKY, ALEX J. (Dr) 4707 Foxhall Cres NW, Washington DC 20007-1064 (EM)

SHAFRIN, ELAINE G. (Mrs.) 8100 Connecticut Ave NW Apt 1014, Washington DC 20815-2817 (EF)

SHROPSHIRE, JR, W. (Dr.) Apt. 426, 300 Westminster Canterbury Dr., Winchester VA 22603 (LF)

SIMMS, JAMES ROBERT (Mr.) 9405 Elizabeth Ct., Fulton MD 20759 (M)

SLUZKI, CARLOS (Dr) 5302 Sherier Pl NW, Washington DC 20016 (F)

SMITH, THOMAS E. (Dr.) 3148 Gracefield Rd Apt 215, Silver Spring MD 20904-5863 (LF)

SNIECKUS, MARY (Ms) 1700, Dublin Dr., Silver Spring MD 20902 (F)

SODERBERG, DAVID L. (Mr.) 403 West Side Dr. Apt. 102, Gaithersburg MD 20878 (EM)

SOLAND, RICHARD M. (Dr.) 2516 Arizona Av Apt 6, Santa Monica CA 90404-1426 (LF)

SOZER, AMANDA (Dr.) 525 Wythe Street, Alexandria VA 22314 (M)

STAVELEY, JUDY (Dr.) 880 Laval Drive, Sykesville MD 21784 (M)

STERN, KURT H. (Dr.) 103 Grant Avenue, Takoma Park MD 20912-4328 (EF)

STIEF, LOUIS J. (Dr.) 332 N St., SW., Washington DC 20024-2904 (EF)

STILES, MARK D. 11506 Taber Street, Silver Spring MD 20902 (F)

STOMBLER, ROBIN (Ms.) Auburn Health Strategies, 3519 South Four Mile Run Dr., Arlington VA 22206 (M)

SUBRAHMANIAN, ESWARAN (Dr.) 4740 Connecticut Avenue, Apt #815, Washington DC 20008 (LM)

SVEDBERG, ERIK (Dr) c/o Washington Academy of Sciences, Suite GL117, 1200 New York Ave, NW, Washington DC 20005 (F)

TEICH, ALBERT H. (Dr.) PO Box 309, Garrett Park MD 20896 (EF)

THEOFANOS, MARY FRANCES (Ms.) 7241 Antares Drive, Gaithersburg MD 20879 (M)

THOMPSON, CHRISTIAN F. (Dr.) 278 Palm Island Way, Ponte Vedra FL 32081 (LF)

TIMASHEV, SVIATOSLAV (SLAVA) A. (Dr.) 3306 Potterton Dr., Falls Church VA 22044-1603 (F)

TORAIN II, DAVID S (Dr.) 1313 Summerfield Drive, Herndon VA 20170 (LM)

TOUWAIDE, ALAIN Botany Center, The Huntington, 1151 Oxford Road, San Marino CA 91108 (LF)

TROXLER, G.W. (Dr.) PO Box 1144, Chincoteague VA 23336-9144 (F)

UMPLEBY, STUART (Professor) Apt 1207, 4141 N Henderson Rd, Arlington VA 22203 (F)

VAISHNAV, MARIANNE P. P.O. Box 2129, Gaithersburg MD 20879 (LF)

VAKHARIA, PRACHI 2400 Virginia Ave. NW, C-1115, Washington DC 20037 (M)

VANE III, RUSSELL RICHARDSON (Dr.) 1713 Pebble Beach Drive, Vienna VA 22182 (M)
VARADI, PETER F. (Dr.) Apartment 1606W, 4620 North Park Avenue, Chevy Chase MD 20815-7507 (EF)
VAVRICK, DANIEL J. (Dr.) 10314 Kupperton Court, Fredricksburg VA 22408 (F)
VILLARUBIA, JOHN 17101 Tom Fox Ave., Poolesville MD 20837 (M)
VOAS, JEFFREY (Dr.) 8210 Crestwood Heights Drive, Apartment 720, McLean VA 22102 (LM)
VOORHEES, ELLEN (Dr.) 100 Bureau Dr., Stop 8940, Gaithersburg MD 20899-8940 (F)
WALDMANN, THOMAS A. (Dr.) 3910 Rickover Road, Silver Spring MD 20902 (F)
WALLER, JOHN D. (Dr.) 5943 Kelley Court, Alexandria VA 22312-3032 (M)
WANG, Y. CLAIRE (Dr.) 140 Charles Street, Apt 22D, New York NY 10014 (M)
WEBB, RALPH E. (Dr.) 21-P Ridge Road, Greenbelt MD 20770 (EF)
WEISS, STEVE (Dr.) 6516 Truman Lane, Falls Church VA 22043 (LF)
WHITE, CARTER (Dr.) 12160 Forest Hill Rd, Waynesboro PA 17268 (EF)
WIESE, WOLFGANG L. (Dr.) 8229 Stone Trail Drive, Bethesda MD 20817 (EF)
WILLIAMS, CARL (Dr.) 2272 Dunster Lane, Potomac MD 29854 (F)
WILLIAMS, E. EUGENE (Dr.) Dept. of Biological Sciences, Salisbury University, 1101 Camden Ave, Salisbury MD 21801 (M)
WILLIAMS, JACK (Mr.) 6022 Hardwick Place, Falls Church VA 22041 (F)

Delegates to the Washington Academy of Sciences Representing Affiliated Scientific Societies

Acoustical Society of America	Paul Arveson
American/International Association of Dental Research	J. Terrell Hoffeld
American Association of Physics Teachers, Chesapeake Section	Frank R. Haig, S. J.
American Astronomical Society	Sethanne Howard
American Fisheries Society	Lee Benaka
American Institute of Aeronautics and Astronautics	David W. Brandt
American Institute of Mining, Metallurgy & Exploration	E. Lee Bray
American Meteorological Society	Vacant
American Nuclear Society	Charles Martin
American Phytopathological Society	Vacant
American Society for Cybernetics	Stuart Umpleby
American Society for Microbiology	Vacant
American Society of Civil Engineers	Vacant
American Society of Mechanical Engineers	Daniel J. Vavrick
American Society of Plant Physiology	Mark Holland
Anthropological Society of Washington	Vacant
ASM International	Toni Marechaux
Association for Women in Science	Jodi Wesemann
Association for Computing Machinery	Vacant
Association for Science, Technology, and Innovation	F. Douglas Witherspoon
Association of Information Technology Professionals	Vacant
Biological Society of Washington	Vacant
Botanical Society of Washington	Chris Puttock
Capital Area Food Protection Association	Keith Lempel
Chemical Society of Washington	Vacant
District of Columbia Institute of Chemists	Vacant
Eastern Sociological Society	Ronald W. Mandersheid
Electrochemical Society	Vacant
Entomological Society of Washington	Vacant
Geological Society of Washington	Jeff Plescia Jurate Landwehr
Historical Society of Washington DC	Vacant
Human Factors and Ergonomics Society	Gerald Krueger

(continued on next page)

Delegates to the Washington Academy of Sciences Representing Affiliated Scientific Societies

(continued from previous page)

Institute of Electrical and Electronics Engineers, Washington Section	Richard Hill
Institute of Food Technologies, Washington DC Section	Taylor Wallace
Institute of Industrial Engineers, National Capital Chapter	Neal F. Schmeidler
International Association for Dental Research, American Section	Christopher Fox
International Society for the Systems Sciences	Vacant
International Society of Automation, Baltimore Washington Section	Richard Sommerfield
Instrument Society of America	Hank Hegner
Marine Technology Society	Jake Sobin
Maryland Native Plant Society	Vacant
Mathematical Association of America, Maryland-District of Columbia-Virginia Section	John Hamman
Medical Society of the District of Columbia	Julian Craig
National Capital Area Skeptics	Vacant
National Capital Astronomers	Jay H. Miller
National Geographic Society	Vacant
Optical Society of America, National Capital Section	Jim Heaney
Pest Science Society of America	Vacant
Philosophical Society of Washington	Larry S. Millstein
Society for Experimental Biology and Medicine	Vacant
Society of American Foresters, National Capital Society	Marilyn Buford
Society of American Military Engineers, Washington DC Post	Vacant
Society of Manufacturing Engineers, Washington DC Chapter	Vacant
Society of Mining, Metallurgy, and Exploration, Inc., Washington DC Section	E. Lee Bray
Soil and Water Conservation Society, National Capital Chapter	Erika Larsen
Technology Transfer Society, Washington Area Chapter	Richard Leshuk
Virginia Native Plant Society, Potowmack Chapter	Alan Ford
Washington DC Chapter of the Institute for Operations Research and the Management Sciences (WINFORMS)	Meagan Pitluck-Schmitt
Washington Evolutionary Systems Society	Vacant
Washington History of Science Club	Albert G. Gluckman
Washington Paint Technology Group	Vacant
Washington Society of Engineers	Alvin Reiner
Washington Society for the History of Medicine	Alain Touwaide
Washington Statistical Society	Michael P. Cohen
World Future Society, National Capital Region Chapter	Jim Honig

Washington Academy of Sciences
Room GL117
1200 New York Ave. NW
Washington, DC 20005
Return Postage Guaranteed